Bayesian Hierarchical Modelling Frameworks for Flawed Data in Environment and Health

Submitted by
Oliver Russell Stoner
to the University of Exeter as a thesis for the degree of
Doctor of Philosophy in Mathematics
In April 2019

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signed: .....

Abstract

In the fields of environment and health, available data is usually not a perfect representation of the quantity we are interested in, such as the number of people contracting a disease or the number of environmental hazards occurring in a given area or time period. Instead, data often suffer from a number of flaws, some of which can pose serious problems. For example, counts of disease cases or environmental hazards may suffer from under-reporting, such that the recorded count is less than or equal to the true count. In some cases, we will never know the true number. This inevitably convolutes our understanding of the risk the disease or natural hazard poses to society. A similar example is delayed reporting of counts, where we may eventually know the true count or something trivially close to it after a period of time. However, we often need to make important decisions, such as how to respond to a disease outbreak, before this certainty is available to us and based on any partial information we may instead have at our disposal.

In this thesis we discuss different ways in which data may be flawed, which we refer to as flawed observation mechanisms, and the risks they pose to practitioners if ignored. Moving beyond previous approaches to tackling this issue, which mostly constitute bespoke solutions to individual problems, we present a conceptual framework for simultaneously modelling quantities we are interested in and any flawed observation mechanisms. We argue that the key strengths of this framework are its ability to rigorously quantify uncertainty, its flexibility and its interpretability. We spend the rest of the thesis demonstrating the power this framework offers to practitioners, with chapters dedicated to the general problems of under-reporting and delayed reporting, as well as a chapter dedicated to the exposition of a model which informs global health policy. Each of these chapters is broadly self contained, with individual discussions of the problems addressed. The thesis concludes with an overview of the effectiveness of our approach and some suggestions for future research.

Supplementary Material

All supplementary material referenced in this thesis can be found at: `https://github.com/orstoner/thesis`.

Publications

As a result of the work presented in this thesis, the following have been published, or are under review (at the time of first submission):

- Stoner, O. (2018). Correcting under-reporting in historical volcano data. *Proceedings of the 33rd International Workshop on Statistical Modelling 1*, 288-292.

- Stoner, O., T. Economou, and G. Drummond Marques da Silva (2019). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association.*

- Stoner, O., G. Shaddick, T. Economou, S. Gumy, J. Lewis, I. Lucio, and H. Adair-Rohani (Under Review). Estimating household air pollution: A multivariate hierarchical model for the use of polluting fuels for cooking.

- Stoner, O., T. Economou (Under Review). Multivariate hierarchical frameworks for modelling delayed reporting in count data.

At the time of final submission, a further article has also been published:

- Stoner, O., G. Shaddick, T. Economou, S. Gumy, J. Lewis, I. Lucio, G. Ruggeri, and H. Adair-Rohani (2019). Multivariate hierarchical modelling of household air pollution. *Proceedings of the 34th International Workshop on Statistical Modelling 2*, 242-247.

Acknowledgments

I begin by thanking my excellent supervisor and friend Theo Economou, whose advice has helped me to grow both as a researcher and as a person. I also thank Chris Ferro for his, at times vital, pastoral support, Gavin Shaddick for helping to build the foundations of my career and all of the other academics in the department whose actions have contributed to where I am today.

I thank all of my friends and family and, finally, I thank my wonderful partner of six years, Susannah Hearn, for her stalwart loyalty, companionship and generosity (including proof-reading this thesis and all of my papers).

I was supported by a NERC GW4+ Doctoral Training Partnership studentship from the Natural Environment Research Council [NE/L002434/1].

Contents

Chapter 1

Introduction

In this chapter we begin by discussing different ways in which data might be flawed and motivate the idea that the mechanisms which cause the flaws should be taken into account in the modelling. We then discuss some ad-hoc solutions for achieving this and their limitations. We also consider an existing conceptual framework for simultaneously modelling a process we're interested in and flawed observation. Building on this, we then propose a modular modelling framework. In this framework, modules which account for one or more flawed observation mechanisms can be added, removed and modified with ease, compared to previous approaches. We then discuss the key strengths of this framework and, finally, the chapter ends with an overview of the content of the thesis.

1.1 Motivation

The use of Bayesian modelling methods is widespread both in the field of environmental statistics (e.g. Economou et al. (2014)), the field of biometric or epidemiological statistics (e.g. Shaddick and Zidek (2015), Shaweno et al. (2017), Broemeling (2013)) and where they interface (e.g. Shaddick et al. (2017)). Modelling practice often involves the use of data which is a flawed representation of processes we are interested in. For example, for natural hazards which rely primarily on direct observation, such as tornados, observed data may be incomplete. This situation can arise due to often unknown mechanisms which occur between the process we are interested in and the modelling of the data.

In many cases, modelling the data without taking these mechanisms into account can be perilous. A well-studied example of a mechanism which presents this issue is under-reporting in count data. This is where the reported counts contained within a dataset, such as of disease cases or natural hazards, are less than or equal to the true counts they are supposed to represent. For counts of disease cases, this situation might arise where healthcare coverage is not universal, where disease notifications systems are inadequately resourced or where sections of the population are difficult

to reach. For counts of natural hazards, meanwhile, under-reporting may arise where reporting relies on human observation, such that reporting is less reliable in areas of low population density.

The danger in modelling count data which may suffer from under-reporting without taking the under-reporting into account is that it leads to biased inference, specifically the under-estimation of the incidence of counts. In the case of disease counts this under-estimation may be most severe in areas with low healthcare coverage, masking the need for intervention. Likewise, failing to take into account under-reporting of natural hazards could lead to the under-estimation of the risk they pose to society. Another consideration is that where under-reporting is a symptom of a wider problem, such as inadequate local healthcare funding, learning about where under-reporting is most severe can be useful to inform improvements in coverage.

In some cases, failing to take into account flawed observation processes may not necessarily lead to a biased inference but instead to a loss of information, which could encumber policy and intervention or even lead to poor decision making. For example, in the reporting of infectious diseases, such as dengue fever, it may take weeks or even months for the number of cases which occurred in a given time period to be near fully reported. In this situation, decision makers may not know whether a severe outbreak has occurred until so much time has passed that an effective intervention is no longer possible. By taking into account the delayed reporting mechanism, including how structures in its severity may vary in space and time, it becomes possible to predict the true count based on any partial counts which have already been observed. In the case of the surveillance of infectious diseases, this can allow for a timely intervention, such as the deployment of additional healthcare resources, which may have been otherwise impossible.

Another example where there is a risk of potential information loss is in household surveys of which fuels people rely on primarily for cooking. While many surveys allow respondents to choose from a comprehensive list of fuels, often some fuels are excluded and are instead relegated to the 'other' category. Similarly, survey choices for fuels with a similar use such as natural gas and liquid petroleum gas are often combined. If, for the sake of modelling, we omit surveys which do not present a comprehensive list, then we are wasting valuable information contained within the partial surveys. This can become problematic if, for example, a given country only has non-comprehensive surveys, leaving us with little to no information on which to base our inference. Instead, we could seek to predict how many people use the fuels which weren't specified individually, based on the partial information contained within the surveys.

In addition, whilst many surveys provide separate data for urban and rural areas, which is important for effective planning of where interventions are most needed, some surveys only present overall data for the whole country. Again, if we can't find

a way of permitting these surveys, so that the urban and rural disaggregation can be inferred, then we are once more wasting valuable information. To further compound matters, some countries may systematically sample too many urban or too many rural people, so that the overall values are biased. As such, the household survey data is also an example of where failing to take into account flawed observation mechanisms can lead to a biased inference.

Flawed observation mechanisms can even interact with each other to make reliable inference even more challenging. For example, infectious disease data may suffer from delayed-reporting but also from under-reporting in the final count, such that the total number of cases reported is still an under-estimate of the true number of cases. For this reason, we require a modular approach where we can simultaneously model the processes we are interested in and take into account any flawed observation mechanisms.

1.2 Background

The overwhelming majority of existing efforts to account for flawed observations are bespoke approaches to solve specific problems. For example, for under-reported, censored or truncated data there exist correction methods based on the censored likelihood (Ibrahim et al. (2001), Lawson (2018)), for delayed reporting there exists the chain-ladder method (Mack, 1993) and for missing values there exist methods based on imputation (Clarke and Hardy (2007), Chapter 7 of Shaddick and Zidek (2015)). The thing that these approaches have in common is that they usually involve separating the task of correcting for the flawed observation mechanisms and the task of modelling the process of interest. For example, Bailey et al. (2005) conduct a study of leprosy cases in Brazil and, before doing any modelling, make a fixed decision that only counts from regions above a chosen poverty threshold are under-reported.

However, these solutions are often restrictive in the sense that they reduce flexibility for other aspects of the model, such as the model for the process we are interested in, or that they can only account for flawed observation mechanisms with simple structures. For example, the chain-ladder method for delayed-reporting assumes that the delay mechanism does not change over time, which is often not sensible when considering time series several years long. Similarly, the censored likelihood method and its derivatives (e.g. Bailey et al. (2005)) do not inherently account for varying levels of severity of under-reporting, which may relate to one or more covariates (Stoner et al., 2019a), and also require that it is known a-priori which data are under-reported. These bespoke solutions may also be restrictive in the sense that they usually don't present any immediate way of coping with more than one flawed observation mechanism. Furthermore, separating the task of

modelling the process of interest and the task of correcting for the under-reporting mechanism means that any joint uncertainty that may exist between them is ignored. For example, when modelling under-reporting we might imagine that our uncertainty in the incidence rate of the true counts is related to our uncertainty in the reporting rate. More examples of ad-hoc solutions and their limitations are discussed specifically for under-reporting and delayed reporting in Chapters 2 and 4, respectively.

To maintain a high level of flexibility, it is more helpful to think in terms of general modelling frameworks for accounting for flawed observation mechanisms. Chapter 8 of Gelman et al. (2014) presents such a conceptual framework, where observed data/information are thought of as a subset of a larger set of complete data, leaving a degree of incompleteness that can manifest as, for example, missing values or censored data. In this framework, both the observed data and the incompleteness mechanism are modelled probabilistically within the Bayesian hierarchical framework. The authors also discuss various incompleteness mechanisms, including those which can be ignored without affecting inference (e.g. random sampling of a finite population), as well as those which must be taken into account (e.g. censoring).

The key advantage of this framework is its rich capacity to quantify uncertainty: set within the Bayesian hierarchical framework, uncertainty in model parameters $\boldsymbol{\theta}$ is rigorously quantified by the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, where $\boldsymbol{y}$ represents any observed data (Shaddick and Zidek, 2015, Chapter 4). Beyond this, posterior predictive model checking makes it possible to check a model's ability to capture any desired aspect of the observed data, and can also be used to perform model selection (Gelman et al., 2014, Chapters 6-7). The flexibility of the Bayesian hierarchical modelling framework allows for the inclusion of both sophisticated covariate relationships and complex spatio-temporal structures, such as the spatial model for tuberculosis incidence in Section 2.3 or the temporal model for dengue fever occurrence in Section 4.5. This makes it possible to use models for purposes such as spatial smoothing or interpolation, prediction of new data and forecasting even in the context of flawed observation mechanisms.

However, for more complicated forms of incomplete data the presentation of this framework in Gelman et al. (2014) becomes strained. For example, in the case of under-reporting, counts which are known to either be under-reported or completely reported can be denoted by a binary indicator variable I, as discussed in more detail in Section 2.2. However, where we wish to quantify the severity of under-reporting, for example by estimating the effects of covariates on the reporting rate, the use of a binary indicator is no longer appropriate (Stoner et al., 2019a). Indeed, Chapter 8 of Gelman et al. (2014) ends with an acknowledgement that, in such cases, the framework must be generalised. While they argue that the Bayesian approach still

holds, this is not expanded upon further.

Building on the strengths of this framework, in the subsequent section we present an alternative way of conceptualising the simultaneous treatment of both processes we are interested in and any flawed observation mechanisms. Treated as modules, the framework allows for an arbitrary number of such mechanisms and interactions between them.

1.3 Modular Framework

The conceptual approach we advocate is to consider the process we are interested in, such as the number of tuberculosis cases per year occurring in a given region, as something that occurs at a latent level within a Bayesian hierarchical framework. Any flawed observation mechanisms are then incorporated in a modular fashion, in between this latent process and the observed data.

As an illustrative example, let y be the number of cases of a disease occurring in a given time period. For this, we have a corresponding model $Y(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are model parameters and/or random effects, such as those related to the rate of cases per year. In the case that there is no under-reporting, we would model observations for y as arising in the following way:

$$Y(\boldsymbol{\theta}) \rightarrow y \tag{1.1}$$

where '$\rightarrow$' means the process $Y(\boldsymbol{\theta})$ generates the quantity y. Now suppose that the cases are under-reported, resulting in observations z such that $z \leq y$. To account for this, we include a further module in the model, $Z(\boldsymbol{\pi})$, depending on parameters and/or random effects $\boldsymbol{\pi}$, such as those relating to the reporting rate of cases. We now consider our observations z as arising from:

$$Y(\boldsymbol{\theta}) \rightarrow y \rightarrow Z(\boldsymbol{\pi}) \rightarrow z \tag{1.2}$$

such that the true number of cases y is unobserved and must be predicted from the model conditional on the observed z. In practical implementations using Markov Chain Monte Carlo (MCMC), these predictions are usually automatically available. In some special cases, we may be able to exploit probability calculus to combine the processes $Y(\boldsymbol{\theta})$ and $Z(\boldsymbol{\pi})$, i.e.:

$$Z(\boldsymbol{\theta}, \boldsymbol{\pi}) \rightarrow z \tag{1.3}$$

which may result in a more efficient implementation. An example of this is given in Section 2.2.2 where the true count generating process and under-reporting mechanism are combined.

So far, we have not departed from the realm of problems which can be addressed by off-the-shelf solutions (e.g. censored likelihood (Lawson, 2018)). However, now suppose that the under-reported cases have also been rounded to the nearest 100

cases, resulting in observations x. In this case, another module can be added to the model:

$$Y(\boldsymbol{\theta}) \rightarrow y \rightarrow Z(\boldsymbol{\pi}) \rightarrow z \rightarrow X(\boldsymbol{\gamma}) \rightarrow x \tag{1.4}$$

While it may be possible with some effort to adjust an off-the-shelf method to take into account this additional flawed observation mechanism, particularly in the simple case of rounding, it is certainly not guaranteed. In contrast, the approach we advocate exploits the flexibility of the Bayesian hierarchical framework, such that modules for different flawed observation mechanisms can be added, removed and modified with ease. This is illustrated in 4.6 where an under-reporting module is added to a framework for modelling delayed reporting. Where flawed observation mechanisms interact with each other, for example if severity of under-reporting and reporting delay are related, we can represent this by modelling the relationships between the parameters of two or more modules (e.g. $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$).

Furthermore, the modular nature of our framework affords a high level of interpretability. This is because the separation of parameters and models associated with each module makes it easier to understand what each one is doing. In the under-reporting example, for instance, random and covariate effects related to the true count are separated from those associated with the under-reporting mechanism, which makes model design, interpretation and the elicitation of prior distributions more straight-forward (Stoner et al., 2019a).

In summary, the key strengths of our approach to simultaneously modelling the processes we are interested in and any flawed observation mechanisms are:

- **Rigorous quantification of uncertainty:** By basing inference on Bayesian (hierarchical) methods, uncertainty in model parameters is fully quantified, while the validity of model assumptions can be assessed using posterior predictive model checking.

- **Flexibility:** The modularity of our approach makes it possible to tackle a variety of problems and even take into account multiple flawed observation mechanisms in a single model.

- **Interpretability:** By presenting the observed data as arising initially from a process we are interested in and then from a series of flawed observation mechanisms, it is easy to understand the role each one plays in generating the observed data.

1.4 Overview

The approach we advocate is quite straight-forward but it is extremely powerful, as we will demonstrate in the remainder of this thesis:

Chapter 2 investigates the specific issue of under-reporting in count data, evaluating previous approaches. We present a general modelling framework for correcting under-reporting in count data, which we thoroughly assess using simulation experiments and illustrate with several applications, spanning both the fields of epidemiology and natural hazards.

Chapter 3 presents a novel Bayesian hierarchical model for household cooking fuel surveys, where changes in the proportion of people using each of 8 key individual fuels are modelled jointly, whilst also incorporating modules in the model to take into account several flawed observation mechanisms, including missing values for some fuels and sampling bias in the proportion of urban and rural respondents.

Chapter 4 investigates the issue of delayed reporting in count data, discussing previous approaches and presenting a general hierarchical framework. The framework, which can also be easily adapted to take into account potential under-reporting of the final observed total count, is tested against previous approaches by means of a case study of reported dengue fever cases in Rio de Janeiro, Brazil.

Finally, Chapter 5 presents some concluding remarks, emphasising the big picture of modelling flawed observation mechanisms and suggesting some potential avenues for future research.

Chapter 2

Under-Reporting of Counts

We begin by noting that the essence of this chapter has been published as a separate article (Stoner et al., 2019a). We present a comprehensive investigation of a Bayesian hierarchical approach to modelling and correcting under-reporting in tuberculosis counts, a general problem arising in observational count data. The framework is applicable to data where all observed counts could potentially be under-reported, relying only on an informative prior distribution for the mean reporting rate to supplement the partial information in the data. Covariates are used to inform both the true count generating process and the under-reporting mechanism, while also allowing for complex spatio-temporal structures. We present several sensitivity analyses based on simulation experiments to aid the elicitation of the prior distribution for the mean reporting rate and decisions relating to the inclusion of covariates. Our principal application of the framework is to tuberculosis data from Brazil, but to highlight its flexibility we also present applications to UK tornado data and global historical volcano data. The chapter ends with a critical evaluation of our approach.

2.1 Introduction

In a variety of fields, such as epidemiology and natural hazards, count data arise which may not be a full representation of the quantity of interest. In many cases the counts are under-reported: the recorded value is less than the true value, sometimes substantially. Quite often, this is due to the observation process being flawed, for instance failing to reach some individuals in a population at risk from infectious disease such as tuberculosis (TB), which is the principal motivating application in this chapter. It is then a missing data challenge and from a statistical point of view, a prediction problem.

The TB surveillance system in Brazil is responsible for detecting disease occurrence and for providing information about its patterns and trends. The notification of TB is mandatory and the data are available in the Notifiable Diseases Information System (SINAN), which provides information about the disease at national, state,

municipal and other regional levels. Despite the high spatial coverage of SINAN, the system is not able to report all TB cases. Using inventory studies (World Health Organization, 2012), the overall TB detection rate for Brazil was estimated as 91%, 84%, and 87% for the years 2012 to 2014 (World Health Organization, 2016).

Under-reporting is an issue because it can lead to biased statistical inference, and therefore poorly informed decisions. This bias will affect parameter estimates, predictions and associated uncertainty. Conventional approaches to quantifying risk, for instance by estimating the spatio-temporal disease rate per unit population, are liable to under-estimate the risk if under-reporting is not allowed for. This has serious societal implications—an estimated 7300 deaths were caused by TB in Brazil in 2016 (World Health Organization, 2016), and this epidemiological burden is masked by under-reporting, which impairs planning of public policies for timely and effective intervention. An alternative system to improve the detection rate has been the active search for cases, especially in high risk groups, including homeless and incarcerated people. However, these activities require local resources, resulting in databases with different detection rates depending on the socio-economic characteristics and the management capacity of the municipalities. It is therefore crucial to estimate and quantify the uncertainty of the detection rates on a finer scale, to allow better informed decisions about the distribution of resources.

In this chapter we investigate a general framework for correcting under-reporting, suitable to a wide range of spatio-temporal count data, and apply it primarily to counts of TB cases in Brazil. To highlight the flexibility of the framework, we also present applications to UK tornado data and historical volcano data, in the field of natural hazards. All counts can be potentially assumed under-reported (unlike other approaches) so that the severity of under-reporting is estimated and potentially informed by available covariates that relate to the under-reporting mechanism. The model is implemented in the Bayesian framework which allows great flexibility and leads to complete predictive distributions for the true counts, therefore quantifying the uncertainty in correcting the under-reporting.

The chapter is structured as follows: Section 2.2 discusses approaches to modelling under-reporting, including the hierarchical framework we will ultimately use, as well as how we seek to resolve the incompleteness of the information provided by the data. Section 2.3 presents the application to Brazilian TB data, as well as some simulation experiments designed to investigate the sensitivity of the model's ability to quantify uncertainty. Sections 2.4 and 2.5 present the application of the framework to UK tornado data and historical volcano data, respectively. Further simulation experiments can be found in the Section 2.6, which address issues such as the sensitivity of the model to the strength of under-reporting covariates. Finally, Section 2.7 presents a critical evaluation of our approach, particularly compared to existing methods.

2.2 Background

Let $y_{i,t,s}$ be the number of events (e.g. TB cases) occurring in units of space $s \in S$, time $t \in T$ and any other grouping structures i that the counts might be aggregated into. If $y_{i,t,s}$ is believed to have been perfectly observed, the counts are conventionally modelled by an appropriate conditional distribution $p(y_{i,t,s} \mid \boldsymbol{\theta})$, usually either Poisson or Negative Binomial. Here $\boldsymbol{\theta}$ represents random effects allowing for various dependency and grouping structures (e.g. space and time), as well as parameters associated with relevant covariates. Inference is then based on the likelihood function (assuming independence in the $y_{i,t,s}$ given $\boldsymbol{\theta}$):

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i,t,s} p(y_{i,t,s} \mid \boldsymbol{\theta}). \tag{2.1}$$

As discussed in Chapter 1, under-reporting is conceptually a form of unintentional missing data (Gelman et al., 2014, Chapter 8) where, in some or potentially all cases, we have not observed the actual number of events $y_{i,t,s}$. Instead, we have observed under-reported counts $z_{i,t,s}$, which represent lower bounds of $y_{i,t,s}$. This implies that using (2.1) for all observed counts, under-reported or otherwise, will lead to biased inference. Rather, we should acknowledge the uncertainty caused by the missing $y_{i,t,s}$, whilst incorporating the partial information provided by the recorded counts $z_{i,t,s}$. More generally, the data collection mechanism should be included in the analysis and this is especially true for missing data problems. A conceptual framework for this (Gelman et al., 2014, Chapter 8) is one where both the completely observed (true) data and the mechanism determining which of them are missing are given probability models. Relating this more specifically to under-reporting, an indicator random variable $I_{i,t,s}$ is introduced, to index the data into fully observed or under-reported. In what follows, we review approaches to under-reporting that can be broadly classified into ones that treat $I_{i,t,s}$ as known, and ones that treat it as latent and therefore attempt to model it.

2.2.1 Censored Likelihood

A common approach to correcting under-reporting is to base inference on the censored likelihood. This is the product of the evaluation of (2.1) for the fully observed (uncensored) counts $y_{i,t,s}$ and the joint probability of the missing $y_{i,t,s}$ exceeding or equalling the recorded (censored) counts $z_{i,t,s}$:

$$p(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\theta}) = \prod_{I_{i,t,s}=1} p(y_{i,t,s} \mid \boldsymbol{\theta}) \prod_{I_{i,t,s}=0} p(y_{i,t,s} \geq z_{i,t,s} \mid \boldsymbol{\theta}). \tag{2.2}$$

In this framework, the indicator $I_{i,t,s}$ for which data are under-reported is binary (where $I_{i,t,s} = 1$ when $z_{i,t,s} = y_{i,t,s}$). The strength of this approach is that all of

the observed counts contribute to the inference and, by accounting for the under-reporting in the model design, a more reliable inference on $\boldsymbol{\theta}$ is possible. However, information on which counts are under-reported is not always readily available, introducing the challenge of having to determine or estimate this classification.

The approach in Bailey et al. (2005) accounts for under-reporting in counts of leprosy cases in the Brazilian region of Olinda, to arrive at a more accurate estimate of leprosy prevalence. They utilise prior knowledge on the relationship between leprosy occurrence rate and a measure of social deprivation to decide the values of $I_{i,t,s}$ a priori: A fixed value of social deprivation is chosen as a threshold, above which observations are deemed to be under-reported. However, the choice of this threshold is subjective and not always obvious. The approach can in principle be extended to include estimation of the threshold, however in many cases the threshold model may be a poor description of the under-reporting mechanism which could, for example, be related to more than one covariate.

Oliveira et al. (2017) presents an alternative to this approach, which treats the binary under-reporting indicator $I_{i,t,s}$ as unobserved and therefore random. The classification of the data is characterised by $I_{i,t,s} \sim \text{Bernoulli}(\pi_{i,t,s})$, such that $\pi_{i,t,s}$ is the probability of any data point suffering from under-reporting, which is potentially informed by covariates. Although a more general approach in the sense of modelling the under-reporting classification, like any other censored likelihood method it lacks a way of quantifying the severity of under-reporting (i.e. the proportion of counts which were not reported and how this relates to covariates). For example, if $\pi_{i,t,s}$ is large, then the corresponding observed count is very likely to suffer from under-reporting, but this tells us nothing about whether the under-reporting is minimal or severe. This makes it unsuitable for our TB application, where we would like to learn about the under-reporting rate on a micro-regional level. Moreover, the predictive inference for the unobserved $y_{i,t,s}$ is limited, amounting to:

$$p(y_{i,t,s} \mid z_{i,t,s}, \boldsymbol{\theta}) = p(y_{i,t,s} \mid y_{i,t,s} \geq z_{i,t,s}, \boldsymbol{\theta}). \tag{2.3}$$

This is because the recorded counts $z_{i,t,s}$ are treated as constants, as opposed to random quantities arising jointly from the $y_{i,t,s}$ process and the under-reporting process, as advocated in the framework presented in Chapter 1. Therefore the severity of under-reporting does not contribute to the predictive inference.

2.2.2 Hierarchical count framework

A potentially more flexible approach is to consider the under-reporting indicator variable $I_{i,t,s}$ as continuous in the range $[0, 1]$, to be interpreted as the proportion of true counts that have been reported. This way, the severity of under-reporting is quantified and estimated when $I_{i,t,s}$ is assumed unknown. One way of achieving this is a hierarchical framework consisting of a Binomial model for the recorded

counts $z_{i,t,s}$ and a latent Poisson model for the true counts $y_{i,t,s}$. This approach, often called the Poisson-Logistic (Winkelmann and Zimmermann, 1993) or Pogit model, has been used across a variety of fields including economics (Winkelmann (2008), Winkelmann (1996)), criminology (Moreno and Girón, 1998), natural hazards (Stoner, 2018) and epidemiology (Greer et al. (2011), Dvorzak and Wagner (2016), Shaweno et al. (2017)). The observed count $z_{i,t,s}$ is assumed a Binomial realisation out of an unobserved total (true) count $y_{i,t,s}$. The basic form of the model (extended in Section 2.3 to include spatial random effects) is given by:

$$
\begin{aligned}
z_{i,t,s} \mid y_{i,t,s} &\sim \text{Binomial}(\pi_{i,t,s}, y_{i,t,s}) && (2.4)\\
\log\left(\frac{\pi_{i,t,s}}{1-\pi_{i,t,s}}\right) &= \beta_0 + \sum_{j=1}^{J} \beta_j w_{i,t,s}^{(j)} && (2.5)\\
y_{i,t,s} &\sim \text{Poisson}(\lambda_{i,t,s}) && (2.6)\\
\log\left(\lambda_{i,t,s}\right) &= \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{i,t,s}^{(k)} && (2.7)
\end{aligned}
$$

All the data can be assumed to be (potentially) under-reported by treating $y_{i,t,s}$ as a latent Poisson variable in a hierarchical Binomial model for $z_{i,t,s}$. Assuming that all individual occurrences have equal chance of being independently reported, $\pi_{i,t,s}$ can be interpreted as the probability that each occurrence is reported, and is effectively the aforementioned indicator variable $I_{i,t,s}$. Relevant under-reporting covariates $\boldsymbol{W} = \{w_{i,t,s}^{(j)}\}$ (e.g. related to TB detection), enter the model through the linear predictor in the logistic transformation of $\pi_{i,t,s}$. This allows inference on the severity of under-reporting and what it relates to.

Whilst in this thesis we generally discuss models for $\pi_{i,t,s}$ which use the logistic link function (as shown in (2.5)) other link functions, such as the complimentary log-log and log links can alternatively be used. For example, use of the log link (under the additional constraint that $\log(\pi_{i,t,s}) \leq 0$) may be desirable where we would like 100% reporting to be achievable in the model.

The true counts $y_{i,t,s}$ are modelled as a latent Poisson variable with mean $\lambda_{i,t,s}$, characterised (at the log-scale) as a linear combination of covariates $\boldsymbol{X} = \{x^{(k)}\}$ associated with the process giving rise to the counts. These are the covariates we would like to capture the effect of, or are known to influence $y_{i,t,s}$, including offsets such as population counts. While here we characterise both the models for $\pi_{i,t,s}$ and $\lambda_{i,t,s}$ as linear combinations of covariates, this is not a restriction of the framework and it is possible to include both non-linear effects and interaction terms. In modelling TB incidence these covariates include social deprivation indicators at a particular location. It is assumed that $\boldsymbol{W}$ and $\boldsymbol{X}$ are comprised of different variables so that the $w_{i,t,s}^{(k)}$ are unrelated to the process generating the counts.

Vectors $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_K)$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_J)$ are parameters to be estimated. Using mean-centred covariates (column means of $\boldsymbol{X}$ and $\boldsymbol{W}$ are zero) implies that

α_0 and β_0 are respectively interpreted as the mean of $y_{i,t,s}$ on the log scale, and the mean reporting rate on the logistic scale, when the covariates are at their means. The framework allows the inclusion of random effects in both (2.5) and (2.7). Random effects allow for overdispersion in count models (Agresti, 2002, Chapter 12), and their inclusion here may be desirable to introduce extra variation and thus flexibility in the model for the true counts, including capturing effects from unobserved covariates. Alternatively, $y_{i,t,s}$ can be NegBin$(\lambda_{i,t,s}, \theta)$: a Negative Binomial with mean $\lambda_{i,t,s}$ and dispersion parameter θ (Winkelmann, 1998). Moreover, some of the coefficients α_k could be assumed random to further increase model flexibility.

We can represent this approach in terms of the modular framework discussed in Section 1.3 as:

$$Y(\lambda_{i,t,s}) \rightarrow y_{i,t,s} \rightarrow Z(\pi_{i,t,s}) \rightarrow z_{i,t,s} \tag{2.8}$$

where we have a latent model $Y(\lambda_{i,t,s})$ for the true counts $y_{i,t,s}$, with a further under-reporting module $Z(\pi_{i,t,s})$ which transforms $y_{i,t,s}$ into the observed counts $z_{i,t,s}$.

Considering the true counts as a latent variable aids in mitigating bias in estimating $\boldsymbol{\alpha}$ from under-reported data. The model is straightforward to implement in the conditional form (2.4)-(2.7), by sampling $y_{i,t,s}$ using Markov Chain Monte Carlo (MCMC). However, doing so will likely result in slow-mixing MCMC chains, owing to high posterior dependence between the samples of $y_{i,t,s}$ and any coefficients or random effects in the model for $\lambda_{i,t,s}$. This means that the chains must be run for a large number of iterations to achieve a desired effective sample size. This can be resolved using the following two results:

Result 1

$$z_{i,t,s} \sim \text{Poisson}(\pi_{i,t,s}\lambda_{i,t,s}) \tag{2.9}$$

Proof:

$$\begin{aligned}
p(z \mid \pi, \lambda) &= \sum_{y=z}^{\infty} p(z \mid y, \pi, \lambda) p(y \mid \lambda) && (2.10)\\
&= \sum_{y=z}^{\infty} \binom{y}{z} \pi^z (1-\pi)^{y-z} \frac{\lambda^y e^{-\lambda}}{y!} && (2.11)\\
&= \sum_{y=z}^{\infty} \frac{y!}{(y-z)!z!} \pi^z (1-\pi)^{y-z} \frac{\lambda^y e^{-\lambda}}{y!} && (2.12)\\
&= \frac{(\pi\lambda)^z e^{-\lambda}}{z!} \sum_{y=z}^{\infty} \frac{(\lambda - \pi\lambda)^{y-z}}{(y-z)!} && (2.13)
\end{aligned}$$

Let $n = y - z$

$$\begin{aligned}
\implies p(z \mid \pi, \lambda) &= \frac{(\pi\lambda)^z e^{-\lambda}}{z!} \sum_{n=0}^{\infty} \frac{(\lambda - \pi\lambda)^n}{(n)!} && (2.14)\\
&= \frac{(\pi\lambda)^z e^{-\lambda}}{z!} e^{\lambda - \pi\lambda} && (2.15)\\
&= \frac{(\pi\lambda)^z e^{-\pi\lambda}}{z!} && (2.16)\\
\implies z \mid \pi, \lambda &\sim \text{Poisson}(\pi\lambda) && (2.17)
\end{aligned}$$

Result 2

$$y_{i,t,s} - z_{i,t,s} \sim \text{Poisson}((1 - \pi_{i,t,s})\lambda_{i,t,s}) \tag{2.18}$$

Proof:

By Bayes' theorem:

$$\begin{aligned}
p(y \mid z, \pi, \lambda) &= \frac{p(z \mid y, \pi, \lambda) p(y \mid \pi, \lambda)}{p(z \mid \pi, \lambda)} && (2.19)\\
&= \frac{\binom{y}{z} \pi^z (1 - \pi)^{y-z} \frac{\lambda^y e^{-\lambda}}{y!}}{\frac{(\pi\lambda)^z e^{-\pi\lambda}}{z!}} && (2.20)\\
&= \frac{\frac{y!}{(y-z)!z!} \pi^z (1 - \pi)^{y-z} \frac{\lambda^y e^{-\lambda}}{y!}}{\frac{(\pi\lambda)^z e^{-\pi\lambda}}{z!}} && (2.21)\\
&= \frac{((1 - \pi)\lambda)^{y-z} e^{-(1-\pi)\lambda}}{(y - z)!} && (2.22)
\end{aligned}$$

This is the Poisson probability mass function for $y - z$ with rate $(1 - \pi)\lambda$, defined on $y \geq z$.

$$\implies y - z \mid z \sim \text{Poisson}((1 - \pi)\lambda) \tag{2.23}$$

Similarly, if $y_{i,t,s} \sim \text{NegBin}(\lambda_{i,t,s}, \theta)$, then $z_{i,t,s} \sim \text{NegBin}(\pi_{i,t,s}\lambda_{i,t,s}, \theta)$. Result 1 is well known, but we have not encountered Result 2 before. The consequence of this is that the model in (2.9) is much more efficient in terms of effective sample size per second, leading to a substantially shorter overall run-time. This is because the $y_{i,t,s}$ no longer need to be sampled during MCMC, greatly alleviating the problem of slow-mixing chains, as well as reducing the number of parameters to sample. Samples of $\boldsymbol{y}$ can be subsequently generated using Monte Carlo simulation of (2.18), meaning that a complete predictive inference on the true counts $y_{i,t,s}$ is possible, deriving information jointly from the mean rate of $y_{i,t,s}$, the reporting probability $\pi_{i,t,s}$ and the recorded counts $z_{i,t,s}$.

However, equation (2.9) suggests that the same observed counts $z_{i,t,s}$ could arise from either a high $\lambda_{i,t,s}$ value combined with a low $\pi_{i,t,s}$, or vice versa, so that the likelihood function of $z_{i,t,s}$ is constant over the level curves of $\pi_{i,t,s}\lambda_{i,t,s}$. This means that, in the absence of any completely reported observations, there is a lack of identifiability between the two intercepts α_0 and β_0. Additionally, as illustrated in

Section 2.6.3, the framework cannot automatically identify whether a given covariate is associated with the under-reporting or the count generating process. This means that care must be taken when deciding which part of the model a covariate belongs in. Non-identifiability for models where the mean is a product of an exponential and logistic term is discussed in greater detail by Papadopoulos and Silva (2012), with discussion more specific to under-reporting in Papadopoulos and Silva (2008).

To conduct meaningful inference on the true counts $y_{i,t,s}$, the partial information in the data must be supplemented with extra information to differentiate between under-reporting and true incidence rate. One potential source of information is to utilise a set of completely reported observations alongside the potentially under-reported observations, an approach used by Dvorzak and Wagner (2016) and Stamey et al. (2006). For these counts, the reporting probability $\pi_{i,t,s}$ (and hence the indicator variable $I_{i,t,s}$) is known a priori to equal 1. In practice, this can be implemented by replacing (2.5) with:

$$\pi_{i,t,s} = c_{i,t,s} + (1 - c_{i,t,s})\frac{\exp(\eta_{i,t,s})}{1 + \exp(\eta_{i,t,s})} \tag{2.24}$$

Here $c_{i,t,s}$ is an indicator variable, where $c_{i,t,s} = 1$ when $z_{i,t,s}$ is completely reported ($\pi_{i,t,s} = 1$) and 0 otherwise ($\pi_{i,t,s}$ is unknown), and $\eta_{i,t,s}$ is the right hand side of (2.5). For some applications, however, such as historical counts of natural hazards (Stoner, 2018), it is often impractical and even impossible to obtain completely observed data. For the application to Brazilian TB data in Section 2.3, complete counts of cases are not available on a micro-regional level. An alternative source of information (Moreno and Girón, 1998) is to employ informative prior distributions to differentiate between $\pi_{i,t,s}$ and $\lambda_{i,t,s}$, which is the approach we adopt in modelling TB. In Section 2.6.1, we examine the effects of either source of information on prediction uncertainty using simulation experiments.

Recently, Shaweno et al. (2017) applied a version of this framework to TB data in Ethiopia, without any data identified as completely observed. However, vague uniform priors are used for regression coefficients, including the intercepts α_0 and β_0. Because of this ambiguity as to whether in practice it is necessary to use an informative prior distribution, we also conduct a thorough investigation of the sensitivity of the framework to the choice of prior distributions using simulated data, in Section 2.3.3.

In summary, the strengths of the hierarchical count framework over the more traditional censored likelihood approach are that it allows both for varying severity of under-reporting across data points and for a more complete predictive inference on the true counts.

2.3 Application to Tuberculosis Data in Brazil

Tuberculosis poses a global health risk and Brazil is among the top twenty countries by absolute mortality. However, this epidemiological burden is masked by under-reporting, which impairs planning for effective intervention. We apply the Bayesian hierarchical approach presented in Section 2.2.2 and in Stoner et al. (2019a) to observed TB counts for the years 2012, 2013 and 2014 and for 557 micro regions. We assume all observed data are potentially under-reported, so we rely on World Health Organization information and simulation experiments to aid in the elicitation of the prior distribution for the mean reporting rate. Covariates and random effects are used to capture variation in both the incidence rate of TB and the reporting probability, including a spatial structure to capture dependency in the TB rate between adjacent micro regions. To assess our model, we present both prior and posterior predictive model checking.

2.3.1 Methodology

Let $y_{t,s}$ and $z_{t,s}$ denote respectively the true and recorded counts of TB cases in micro region $s \in \{1, \ldots, 557\}$ (spanning all of Brazil), and year $t \in \{2012, 2013, 2014\}$. Figure 2.1 illustrates the recorded TB incidence rate per 100,000 people for each year. Whilst there is some variation between the years, the characteristics of the spatial distributions appear generally similar, with large areas of low incidence in the south of Brazil and just south of the centre. This is in contrast to higher incidence in the north-west.

Some of this variability may be attributed to spatial covariates affecting TB incidence. In particular, high risk populations include poorly integrated groups due to poverty related issues, such as homelessness and incarceration. To allow for this, various social deprivation indicators for each micro-region were considered as covariates. These were: $x_s^{(1)}$ = unemployment (the proportion of economically active adults without employment); $x_s^{(2)}$ = urbanisation (the proportion of people living in an urban setting); $x_s^{(3)}$ = density (the mean number of people living per room in a dwelling); and $x_s^{(4)}$ = indigenous (the proportion of the population made up by indigenous groups). These covariates are derived from the 2010 demographic census of Brazil and as such are not temporally indexed.

Figure 2.2 shows scatter plots of these four covariates against the observed TB rate per 100,000 inhabitants. These plots suggest positive relationships between each of the four covariates and the TB rate, though none of these are a clear linear relationship. For example, in the top left panel, it looks like unemployment ($x_s^{(1)}$) may have a positive correlation with the TB rate in the bulk of values, but some of the highest TB observations are for relatively low unemployment values. The effect of density ($x_s^{(3)}$) on the TB rate is also not clear, as on the surface there appears

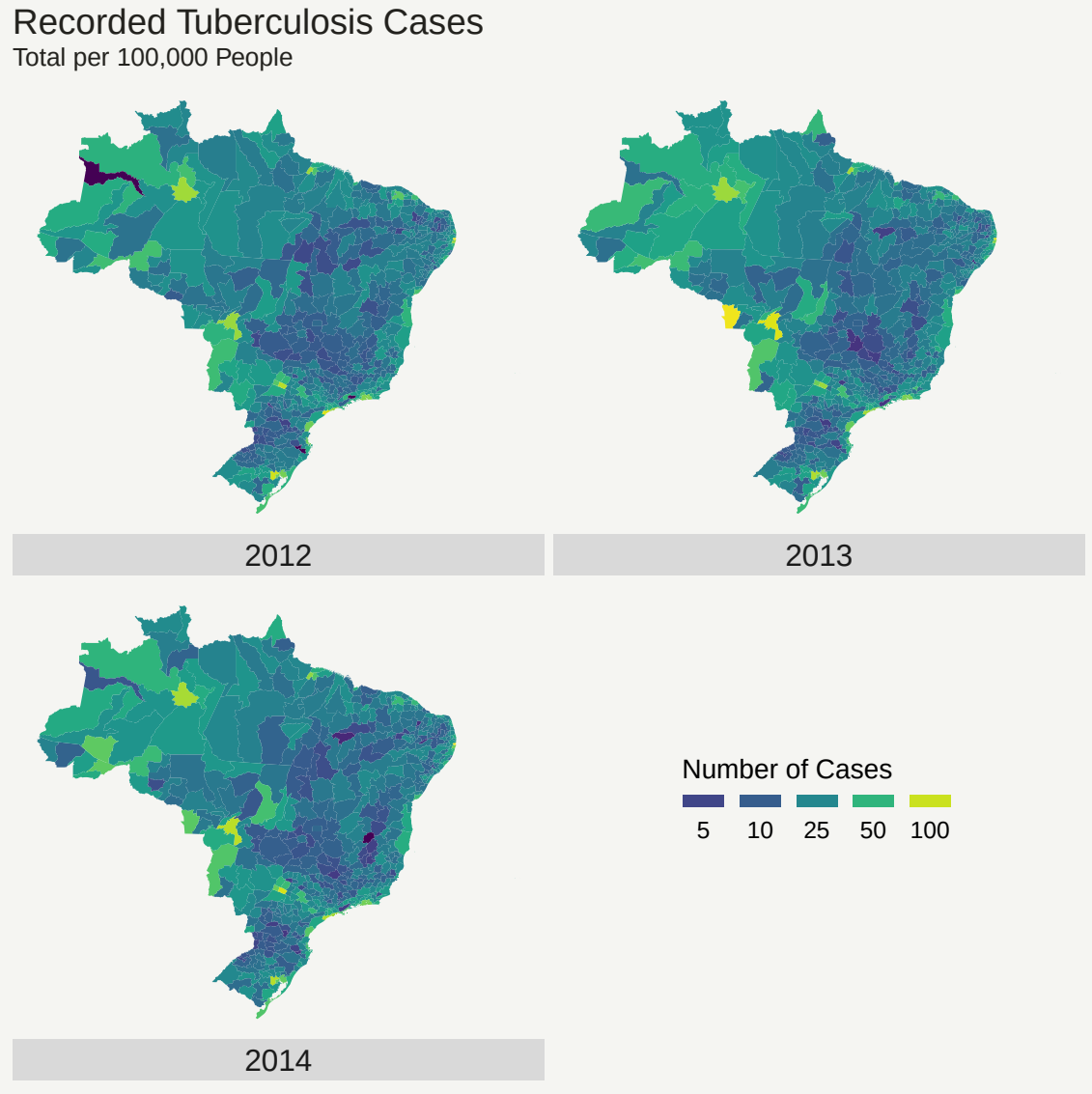

Figure 2.1: Number of observed new TB cases for each mainland micro region of Brazil, for the years 2012-2014, per 100,000 inhabitants.

to be some kind of combination of a steep positive relationship and a more shallow one. Furthermore, there is little evidence of a relationship between urbanisation ($x_s^{(2)}$) and the TB rate, except for in the upper end of urbanisation values where the relationship appears non-linear (with increasing gradient). If they exist, the lack of evidence for clear relationships in these plots may be because we are only looking at the marginal relationships between each covariate and the TB rate. Another possibility is that these relationships are masked by changing demographics since the 2010 census, spatial-clustering of TB incidence, or even the under-reporting mechanism.

Whilst these covariates were considered for inclusion in the model for TB incidence, the covariate u_s = treatment timeliness (the proportion of TB cases for which treatment begins within one day) was considered in the characterisation of the under-reporting mechanism. Having already controlled for social deprivation

Figure 2.2: Scatter plots comparing unemployment, urbanisation, dwelling density, and indigenous proportion covariates to the observed tuberculosis rates, per 100,000 inhabitants.

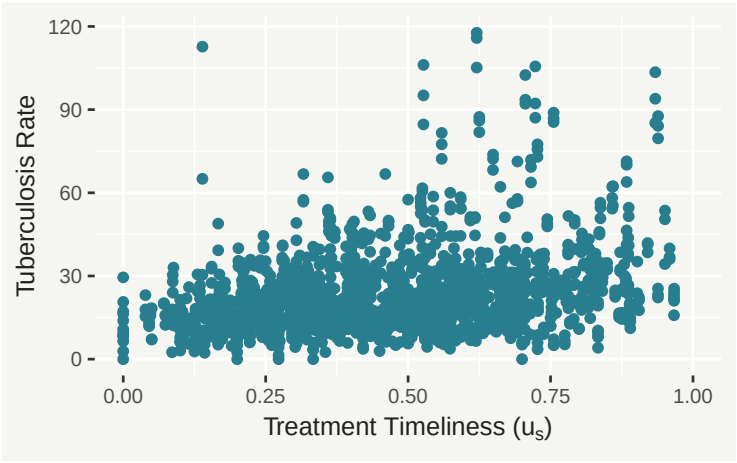

Figure 2.3: Scatter plot of treatment timeliness against the observed tuberculosis rate, per 100,000 people.

through $x_s^{(j)}$, u_s acts as a proxy for how well a local TB surveillance programme is resourced. This implies there should be a positive relationship between timeliness and the TB rate, which is apparent in the scatter plot shown in Figure 2.3. The

model is specified (conditionally on random effects) as follows:

$$
\begin{aligned}
z_{t,s} \mid y_{t,s}, \gamma_{t,s} &\sim \text{Binomial}\left(\pi_s, y_{t,s}\right) && (2.25)\\
\log\left(\frac{\pi_s}{1-\pi_s}\right) &= \beta_0 + g(u_s) + \gamma_{t,s} && (2.26)\\
y_{t,s}|\phi_s, \theta_s &\sim \text{Poisson}(\lambda_{t,s}) && (2.27)\\
\log\left(\lambda_{t,s}\right) &= \log(P_{t,s}) + \alpha_0 + f_1(x_s^{(1)}) + f_2(x_s^{(2)}) \\
&+ f_3(x_s^{(3)}) + f_4(x_s^{(4)}) + \phi_s + \theta_s && (2.28)\\
\phi_s &\sim \text{ICAR}(\nu^2) && (2.29)
\end{aligned}
$$

Functions $g(\cdot)$, $f_1(\cdot), \ldots, f_4(\cdot)$ are orthogonal polynomials of degrees 3, 2, 2, 2 and 1, respectively. Compared to raw polynomials, these reduce multiple-collinearity between the monomial terms (Kennedy and Gentle, 1980), and were set up using the "poly" function in R (R Core Team, 2018). The polynomials are defined such that $f(x) = 0$ when $x = \bar{x}$, so that (at the logistic scale) β_0 is the mean reporting rate for a region with mean treatment timeliness. The term $\log(P_{t,s})$, where $P_{t,s}$ is population, is an offset to allow for varying population and ensure the covariates act on the incidence rate.

Additive effects from a spatially unstructured random effect θ_s and a spatially structured one, ϕ_s are assumed to capture any residual spatial variation in the incidence of TB. An Intrinsic Gaussian Conditional Autoregressive (ICAR) model (Besag et al., 1991) was assumed for ϕ_s, with variance parameter ν^2, to capture dependence between neighbouring micro-regions. Here, a neighbour of s was defined as any $s' \neq s$ sharing a geographical boundary with s. The ICAR model is a reasonable choice here as it is computationally convenient in an MCMC implementation and it is not too smooth. This is important given that we are interested in capturing both any geographical effects on TB incidence and any unobserved spatially-varying covariates. The $\text{N}(0, \sigma^2)$ effect θ_s was included to afford extra spatial residual variability. We can assess our choice of spatial model later by examining the relative dominance of the structured effect ϕ_s and the unstructured effect θ_s. Specifically, if the posterior variance of θ_s dominates the variance of ϕ_s then we might conclude that the ICAR model is either ineffective or inappropriate in this situation. An additional unstructured $\text{N}(0, \epsilon^2)$ effect $\gamma_{t,s}$ was included in the model for the reporting rate (2.26), to allow for the effect of potential unobserved covariates on the detection rate of TB, as well as the case that u_s may only be a proxy for the appropriate (true) under-reporting covariate.

The prior distribution for α_0 was assumed $\text{N}(-8, 1)$, chosen by using prior predictive checking to reflect our belief that very high values (such as over 1 million) for the total number of cases are unlikely. The priors for α_j $(j = 1, ..., 7)$ and β_k $(k = 1, 2, 3)$, the coefficients of the orthogonal polynomials, were specified as $\text{N}(0, 10^2)$, which were chosen to be relatively non-informative. Finally, the priors for

variance parameters σ, ν and ϵ were specified as zero-truncated $N(0, 1)$, to reflect the belief that low variance values are more likely than higher ones, but that these effects are likely to capture at least some of the variance. As discussed in Section 2.2.2, in the absence of any completely reported TB counts, we must specify an informative prior distribution for β_0 to supplement the partial information in the data. As an aid in doing so, we investigate the sensitivity of the model to this prior through simulation experiments presented in the following subsection.

2.3.2 Implementation with NIMBLE

All models in this thesis were implemented using NIMBLE (de Valpine et al., 2017), a facility for flexible implementations of MCMC models in conjunction with R (R Core Team, 2018).Models are written in the BUGS language (like JAGS (Plummer, 2003) and OpenBUGS/WinBUGS (Lunn et al., 2009)) and then compiled automatically for fast execution in C++. The most commonly used distributions (e.g. Poisson) are included by default, but with NIMBLE it's straight-forward to add new distributions, as well as user-defined functions, which are written in R and also compiled. By default, NIMBLE uses a combination of univariate Metropolis-Hastings random walk sampling algorithms, multivariate random walks and slice samplers (exploiting conjugate relationships where possible), though users can manually assign different samplers to single parameters or groups of parameters as desired. New sampling algorithms can also be added in the same way as user-defined distributions and functions. This flexibility is the reason we opted for NIMBLE over other alternatives, such as JAGS. NIMBLE also includes a variety of uncommon MCMC sampling algorithms which can be very effective in some situations.

This work did not involve any explicit attempt to find the most optimal sampling algorithms for each model (in terms of metrics such as the effective sample size per second) and, as such, we generally relied on the default random walk, blocked random walk and slice samplers. Where MCMC performance was particularly poor, for example where the chains were slow mixing, non-standard samplers were adopted and retained if they led to a substantial improvement. For example, in this application we made use of the automated factor slice sampler (AFSS) which can be an efficient way of sampling vectors of highly correlated parameters (Tibbits et al., 2014), such as α_0 and β_0. The associated code and data are provided in the Supplementary Material.

2.3.3 Simulation experiments

As all of the data considered in this chapter are assumed to be entirely susceptible to under-reporting, we rely on simulated data to construct an controlled environment in which we can illustrate and test the effectiveness of our approach. Initially, we

would like to assess the sensitivity of our approach to the choice of prior for β_0, the overall reporting rate (at the logistic level) when all covariates are at their mean. For this experiment, we consider counts which vary in space in the following way:

$$
\begin{aligned}
z_s|y_s &\sim \text{Binomial}(\pi_s, y_s) && (2.30)\\
\log\left(\frac{\pi_s}{1-\pi_s}\right) &= \beta_0 + \beta_1 w_s && (2.31)\\
y_s|\phi_s &\sim \text{Poisson}(\lambda_s) && (2.32)\\
\log(\lambda_s) &= \alpha_0 + \alpha_1 x_s + \phi_s && (2.33)
\end{aligned}
$$

with $\beta_0 = 0$, $\beta_1 = 2$, $\alpha_0 = 4$, $\alpha_1 = 1$ and $\nu = 0.5$. A total of $s = 1, \ldots, 100$ data points were simulated with both covariates x_s and w_s being sampled from a Unif$(-1, 1)$ distribution. The choice of these parameter values is arbitrary, but was made to produce clear (but not extreme) covariate effects. The ICAR(ν^2) spatial effect ϕ_s was simulated over a regular 10x10 lattice. Figure 2.4 shows the simulated data. Note there are clear positive relationships between x_s and y_s, and between w_s and z_s, while there is no clear relationship between w_s and y_s. We proceed to investigate the sensitivity of the model to the specification of the Gaussian prior distribution for β_0, by repeatedly applying the model whilst varying the mean and standard deviation for this prior. The prior for α_0 was $\text{N}(0, 10^2)$, with all other priors the same as in the TB model.

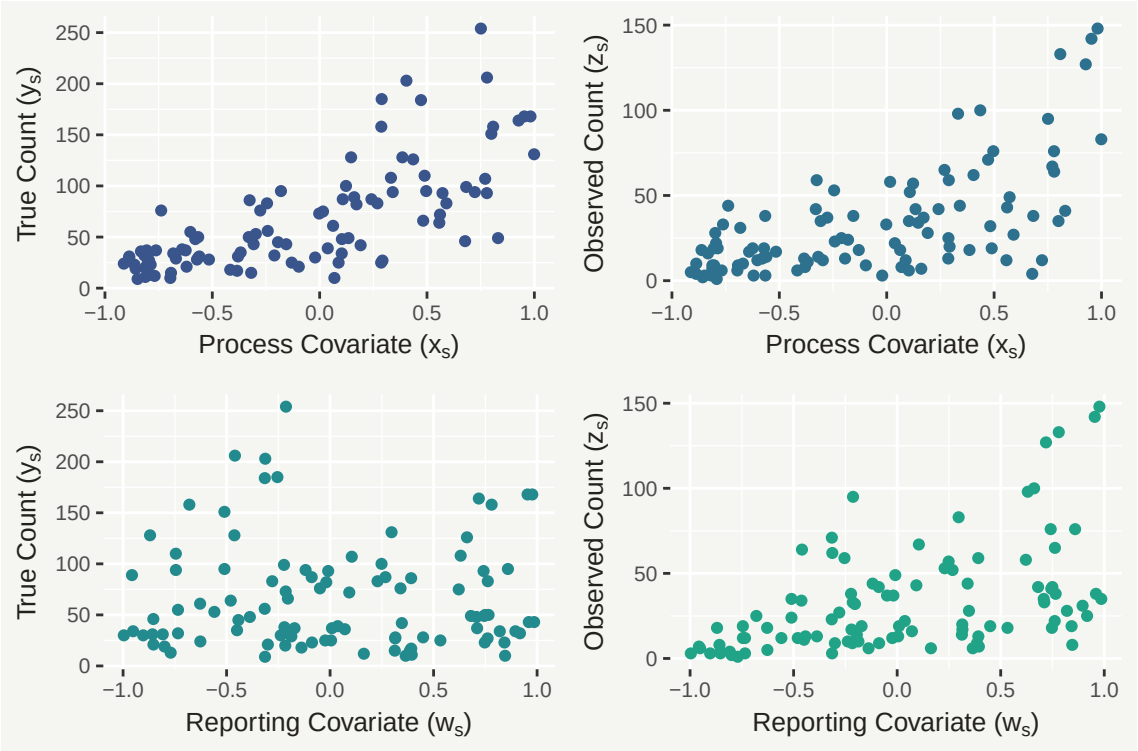

Figure 2.4: Scatter plots of simulated data, showing the process covariate x_s (top row) and the under-reporting covariate w_s (bottom row) against the true counts y_s (left column) and the recorded counts z_s (right column).

To make the experiment more realistic, we mimic the case where the true under-reporting covariate w_s is not available, and instead we only have access to (proxy) covariates $v_{s,2}, ..., v_{s,6}$. These are simulated such that they have decreasing correlation with w_s. As the variation in π_s is no longer fully captured by $v_{s,2}, ..., v_{s,6}$, we include a random quantity $\gamma_s \sim \mathrm{N}(0, \epsilon^2)$ in (2.31).

An important aspect of model performance to consider is the proportion of true counts that lie in their corresponding 95% posterior prediction intervals (PIs), known as the coverage. In the context of non-identifiability, we would expect the coverage to remain high as long as the true value of β_0 is not extreme with respect to its prior. Figure 2.5 shows the coverage when the covariate $v_{s,3}$ (correlation 0.6 with w_s) is used (which incidentally has a similar correlation value with the recorded counts as treatmeant timeliness in the TB data). The plot suggests that the model is able to quantify uncertainty well, as long as a strong prior distribution is not specified well away from the true value (lower corners). The inclusion of γ_s implies that using a "weaker" under-reporting covariate should have little impact on coverage (the PIs of y_s would simply widen). Indeed, more detailed results in Section 2.6.2 show that mean coverage did not change systematically when weakening the covariate.

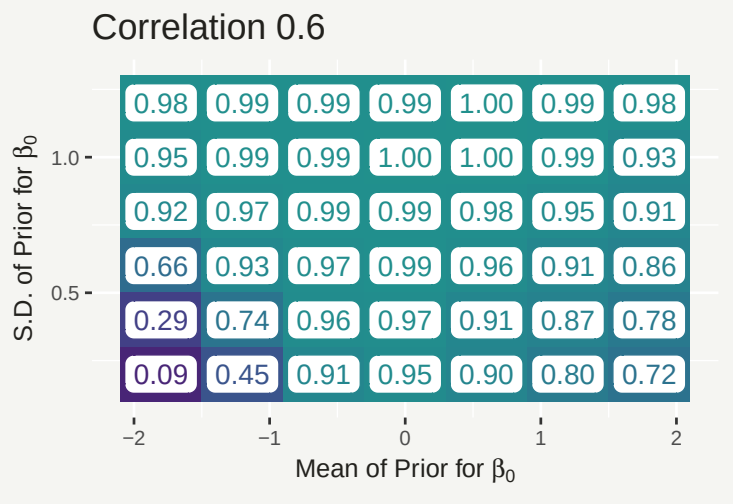

Figure 2.5: Coverage of the 95% PIs for y_s, when the under-reporting covariate $v_{s,3}$, which has a theoretical correlation of 0.6 with the true covariate w_s, is used. Note the true value of β_0 is 0.

As an illustrative example of model performance, Figure 2.6 shows various results based on simulated data using $v_{s,3}$ as the under-reporting covariate, and a $\mathrm{N}(0.6, 0.6^2)$ prior for β_0. This represents the case where the prior distribution overestimates the reporting probability but not to an extreme extent. The top left and middle-left plots show posterior densities for α_0 and α_1, indicating substantial learning of these parameters compared to the flat priors also shown. The top right plot compares the mean predicted spatial effects to their corresponding true values, suggesting these are captured well. The lower-left plot shows the posterior for β_0 has shifted in the direction of the true value. This illustrates that, at least in this idealised setting, the model is not entirely at the mercy of the accuracy of this prior, despite non-identifiability. The middle-right plot shows the mean predicted effect of the imperfect covariate $v_{s,3}$ on the reporting probability, with associated 95% cred-

ible interval (CrI). The effect is quite uncertain, reflecting the relative weakness of the covariate. Finally, the lower right plot shows the lower (blue) and upper (green) limits of the 95% PIs for y_s, suggesting that the model is able to systematically predict well the true unobserved counts.

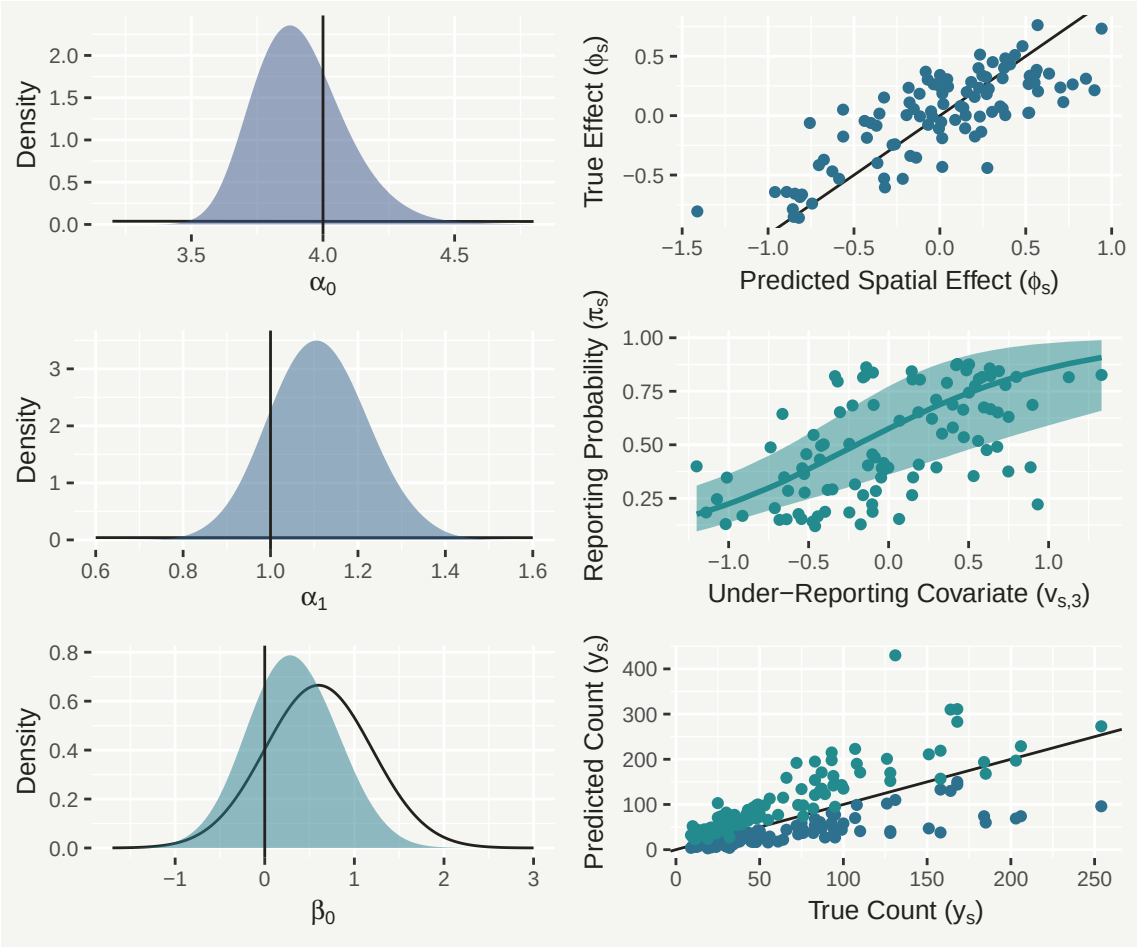

Figure 2.6: Density estimates (left row) of prior (black) and posterior (coloured) samples for parameters α_0, α_1 and β_0, respectively, with vertical lines representing their true values. The top-right plot shows the mean predicted spatial effect (ϕ_s) against the true values. The middle-right plot shows the predicted relationship (solid line) between the under-reporting covariate $v_{s,3}$ and the reporting probability π_s, with associated 95% CrI. The lower-right plot shows the lower (blue) and upper (green) limits of the 95% PIs for the true counts y_s.

This sensitivity analysis is by no means exhaustive, but it does appear to suggest that the model with no completely observed values is robust in terms of quantifying uncertainty, as long as the practitioner specifies a prior for β_0 that is informative but not too strong. With this in mind, we return to the task of specifying this prior distribution for the TB model. The information available are WHO inventory study-derived estimates (World Health Organization, 2012) of the overall TB detection rate in Brazil for 2012-2014. The 2017 point estimates for these years, with associated 95% confidence intervals were 91% (78%,100%), 84% (73%,99%) and 87%

(75%,100%) (World Health Organization, 2016). Normal distributions were used to approximate each rate at the logistic level. We inferred mean and standard deviation parameter values by attempting to match the quoted point estimates and confidence intervals. The mean of the three rates is most variable when they are positively correlated, so to account for this we simulated and sorted into ascending order samples from each approximate distribution, before computing the mean of each sample of three rates. This resulted in a distribution which was approximately $N(2, 0.4^2)$. Figure 2.5 suggests that the mean of this prior can only be slightly wrong (less than 0.5 away) before coverage begins to drop below ideal levels (95%). For this reason, and because the incorporation of the WHO uncertainty is only approximate, we opt for a more conservative standard deviation of 0.6, which allows the mean to deviate more from the truth before PIs become less trustworthy.

2.3.4 Model checking

When using MCMC to implement models, care should be taken to ensure the chains have converged to the target (posterior) distributions. In the first instance, this can be assessed by examining trace plots. These are constructed by plotting the samples of a parameter as a line, for each chain. Figure 2.7 shows trace plots for

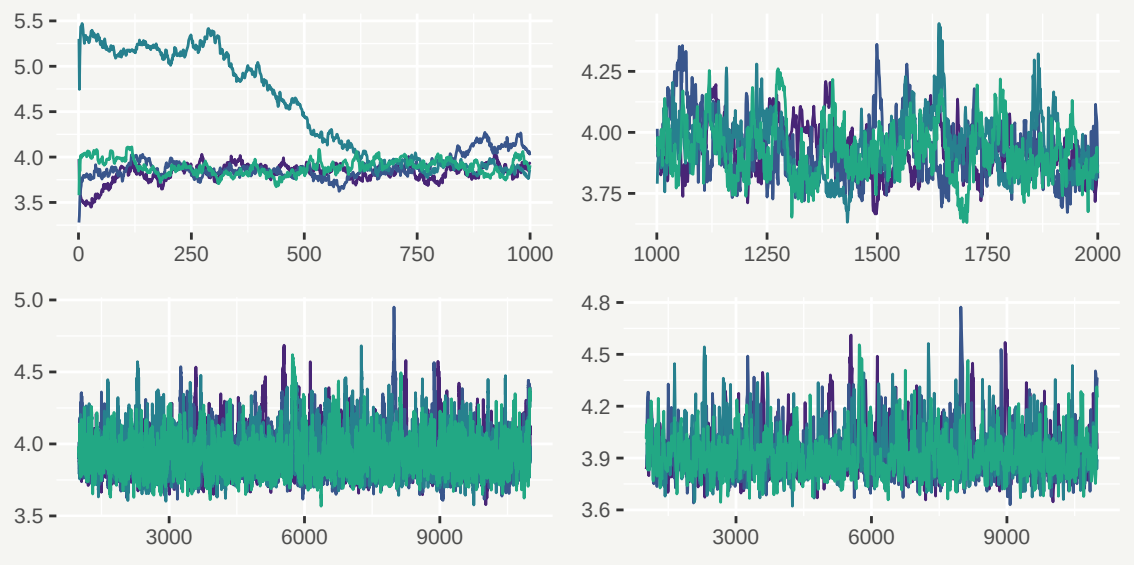

Figure 2.7: Posterior trace plots for the parameter α_0, from the simulation experiment described in Section 2.6.1. Samples 1 to 1000, 1001 to 2000, and 1001 to 11000 are shown in the top-left, top-right, and bottom-left plots, respectively. The bottom-right plot shows every 10^{th} sample from 1010 to 11000.

the parameter α_0, from the simulation experiment described in Section 2.6.1. In the top-left plot, we can see the first 1000 samples from each of the four chains. Each chain starts at a different value and we can see that, while three of the chains appear to have converged within the first 200 iterations or so, one of the chains took

a lot longer to join them. If we include this period before convergence, known as 'burn-in', the distribution of the resulting samples will be severely skewed. For this reason, we may choose to discard the first 1000 samples in this particular example.

The top-right plot then shows samples 1001 to 2000. The chains all appear to be sampling within the same area of the parameter space, suggesting they have likely converged, so discarding the first 1000 samples was probably sufficient here. However, we can see that the chains are quite strongly auto-correlated (in other words they are 'slow mixing'). This means that sample t has a strong correlation with sample $t-1$, within a given chain. While we have 1000 samples for each chain, the 'effective sample size' is lower (Gelman et al., 2014). This means that these samples are in some sense 'lower quality' than samples simulated directly from the posterior. In practice this translates to less reliable point estimates and prediction intervals.

To compensate for this, and to achieve a desired effective sample size, we require an increased number of MCMC samples. The bottom-left plot shows samples 1001 to 11000. However, in some cases system memory limits the number of samples we can store. For example, in the model for household air pollution presented in Chapter 3, we have tens of thousands of parameters to save samples of. If we save a large number of samples for each parameter (e.g. 100k), we will quickly exceed the memory available in a modern desktop computer. We may instead choose to 'thin' these samples. Thinning by n involves saving only every n^{th} sample. As the distance between each saved sample (in terms of the number of iterations) is increased, the auto-correlation between the samples is reduced. The result is that the saved samples have a higher effective sample size than the same number of samples with no thinning. The bottom-right plot shows every 10^{th} sample from 1010 to 11000, i.e. the samples shown in the bottom-left plot but thinned by 10.

As well as inspecting trace plots, we can assess convergence by computing the potential scale reduction factor (PSRF) for each parameter (Brooks and Gelman, 1998), which compares the between-chain and within-chain variances. If the chains have not converged, the between-chain variance should exceed the within-chain variance and the PSRF will be substantially greater than 1. Using different initial values and random number seeds for each chain gives the best assurance that the chains have converged to the whole posterior, rather than a local mode. Four chains were used, each ran for a total of 800K iterations. After discarding 400K iterations as burn-in, the PSRF was computed as less than 1.05 for all regression coefficients and variance parameters. Here and elsewhere we follow the convention that a PSRF of less than 1.05 is sufficiently close to 1 to indicate convergence. This has no strict theoretical justification, but has been widely adopted by the community.

Throughout this thesis, we will fit several models using MCMC and will adopt different values for the number of iterations, the amount to discard as burn-in and the

thinning interval. These values vary substantially from model to model, and broadly reflect different mixing characteristics (slower-mixing models required a higher number of iterations), how many iterations each model took to convergence and system memory limitations. They were chosen roughly such that there was no evidence the chains had not converged in the trace plots, and to satisfy the criteria that the PSRF does not exceed 1.05.

A natural way of assessing whether the model fits the data well is to conduct posterior predictive model checking (Gelman et al., 2014, Chapter 6). More specifically, one can look at the discrepancy between the data $\boldsymbol{z}$ and posterior predictive replicates of this data from the fitted model. Define the posterior predictive distribution for a replicate $\tilde{z}_{t,s}$, of observed number of TB cases $z_{t,s}$, as $p(\tilde{z}_{t,s} \mid \boldsymbol{z})$. The question is then whether the actual observation $z_{t,s}$ is an extreme value with respect to $p(\tilde{z}_{t,s} \mid \boldsymbol{z})$ and if so, this indicates poor model performance.

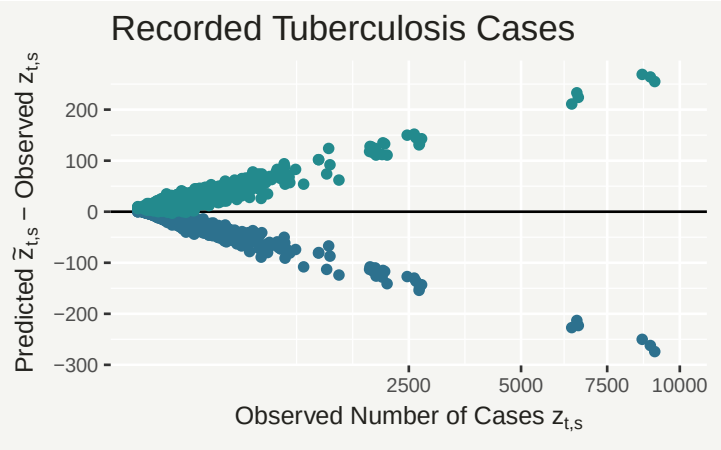

Figure 2.8: Scatter plot of differences between the lower (blue) and upper (green) limits of the 95% PIs of $\tilde{z}_{t,s}$ and the observed values $z_{t,s}$.

Figure 2.8 shows a scatter plot of the difference between the lower (blue) and upper (green) limits of the 95% posterior PIs of $\tilde{z}_{t,s}$ and the corresponding observed values $z_{t,s}$. The PIs are symmetrically centred on the observed values, suggesting that the model has no systematic issue (under or over-prediction) with fitting observed values. The coverage of the 95% PIs was approximately 99.6%.

Furthermore, we can assess whether summary statistics of the original data are captured well by the model through the replicates. Given this is count data, we want to ensure that both the sample mean and variance are captured well. As the prior distributions used for regression coefficients were quite broad, it is important to also assess whether substantial learning has occurred, with respect to both the predictive error of the observed counts $z_{t,s}$ and the distributions of these statistics. Otherwise, it is possible that the data are well captured in the posterior predictions because they were contained within the prior predictions.

The top and middle rows of Figure 2.9 show the prior (left) and posterior (right) predictive distributions of the sample mean and variance. The corresponding observed quantities are in the bulk suggesting that the prior and posterior models capture these well. The posterior predictive distributions are far more precise, in-

dicating that the uncertainty in the parameters has been reduced significantly by the data. This is emphasised by the bottom row, which compares the posterior and prior predictive distributions of the mean squared difference between each $\tilde{z}_{t,s}$ and $z_{t,s}$. The mean squared error is several orders of magnitude smaller in the posterior model, implying far greater prediction accuracy.

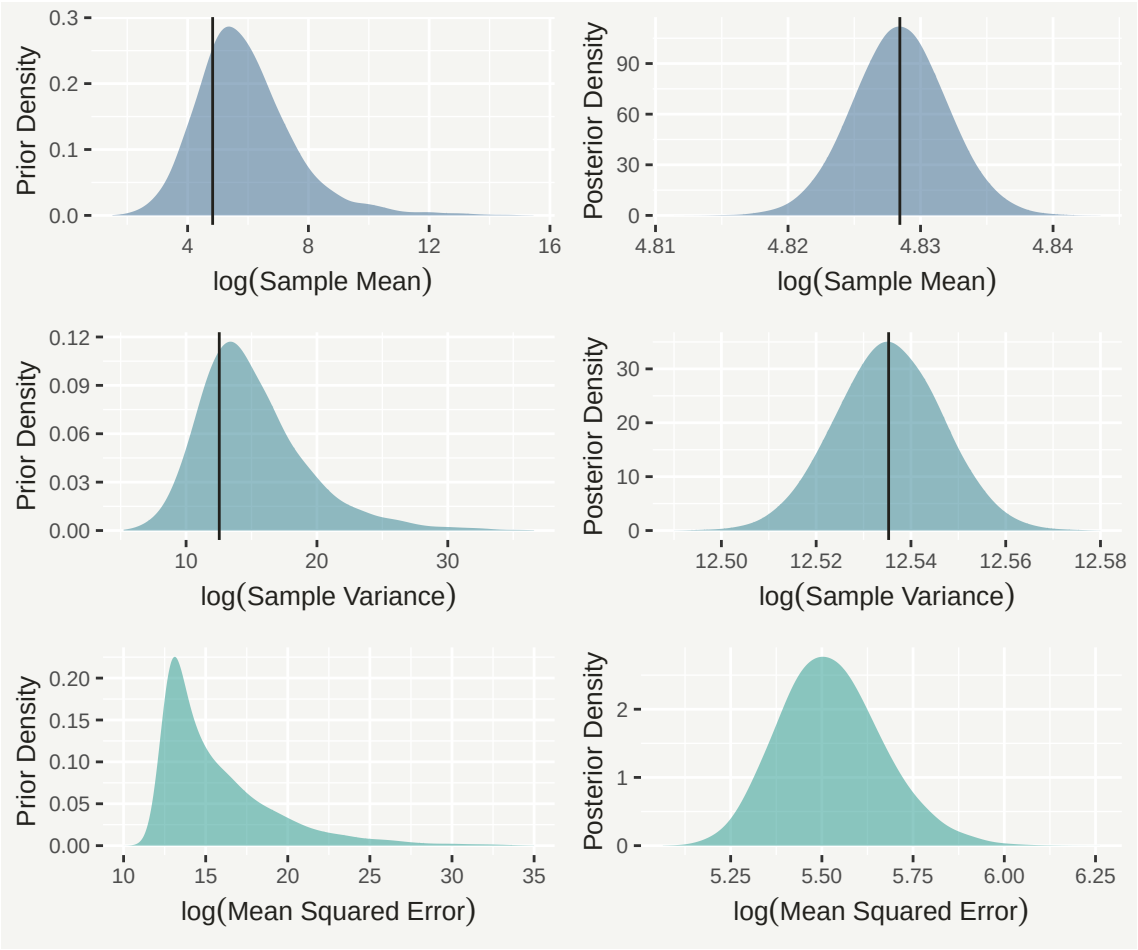

Figure 2.9: Prior (left column) and posterior (right column) predictive distributions of the sample mean (top row), sample variance (middle row) and the log-mean squared error from the recorded counts $z_{i,t,s}$ (bottom row), of the replicates $\tilde{z}_{t,s}$. Observed statistics are plotted as vertical lines.

2.3.5 Results

The effect of unemployment on $\lambda_{t,s}$ is shown in the upper-left panel of Figure 2.10, indicating a strong (based on the width of the 95% CrIs) positive relationship with TB incidence. This is likely because areas with high unemployment often also have high rates of homelessness and incarceration, two important risk factors for TB. The range of this effect is approximately 0.8 on the log scale, suggesting incidence rate is over twice as high in micro-regions with high unemployment ($> 15\%$), compared to areas with low unemployment ($< 5\%$). The lower-left panel shows that urbanised proportion is also strongly positively related to TB incidence. The range of this

effect is also approximately 0.8, meaning that highly urbanised ($> 90\%$) micro-regions are predicted to have over double the TB incidence of micro-regions with low urbanisation ($< 40\%$). This could be due to the increased population density of highly urbanised areas, which may promote the spread of the disease. The effect of dwelling density is less pronounced: the polynomial increases monotonically for most of the range covered by the data ($x_s^{(3)} < 1$), before decreasing for higher values. This suggests that TB incidence is actually lower in micro-regions with the highest levels of dwelling density. Alternatively it may be that further under-reporting of TB is present in such areas, which is not being captured by this model. Data at these upper values are quite sparse, as reflected by widening of the 95% CrIs. Finally, the lower-right panel of Figure 2.10 shows the effect of indigenous proportion. Recall that this relationship was constrained to be linear in (2.28) and the 95% CrI on the slope suggests the effect is strongly positive.

Figure 2.10: Posterior mean predictions (solid lines) of the effects of unemployment, indigenous, density and urbanisation on the rate of TB incidence, with associated 95% CrIs.

Figure 2.11 illustrates the predicted residual spatial variability in the TB incidence rate ($\phi_s + \theta_s$). There is substantial clustering of negative values in the centre of Brazil, surrounding the states of Goiás and Tocantins, while there is clustering of positive values in the North West, including the Amazon rainforest. Interestingly, this seems to align well with estimates of the spatial distribution of human development index (HDI) (see for instance Atlas (2013)), where high estimates of HDI coincide with low values from the spatial effect. This could indicate that there exist other effects of human development on TB incidence, such as healthcare infrastructure, which are not captured by the covariates. Several big cities, including Rio de

Janeiro and São Paulo appear to buck this trend, with positive spatial effects despite relatively high HDI estimates, which could be due to the effect of features unique to big cities, such as high population density, which aren't included in the model. The effect of the spatially structured ϕ_s is visible by the clustering of similar colours and we found it dominated the unstructured effect θ_s, explaining a predicted 94% of their combined variation. The range of values of the combined effect is not dissimilar to the effects of any of the individual covariates, implying that the covariates are driving most of the variability in the true counts $y_{t,s}$.

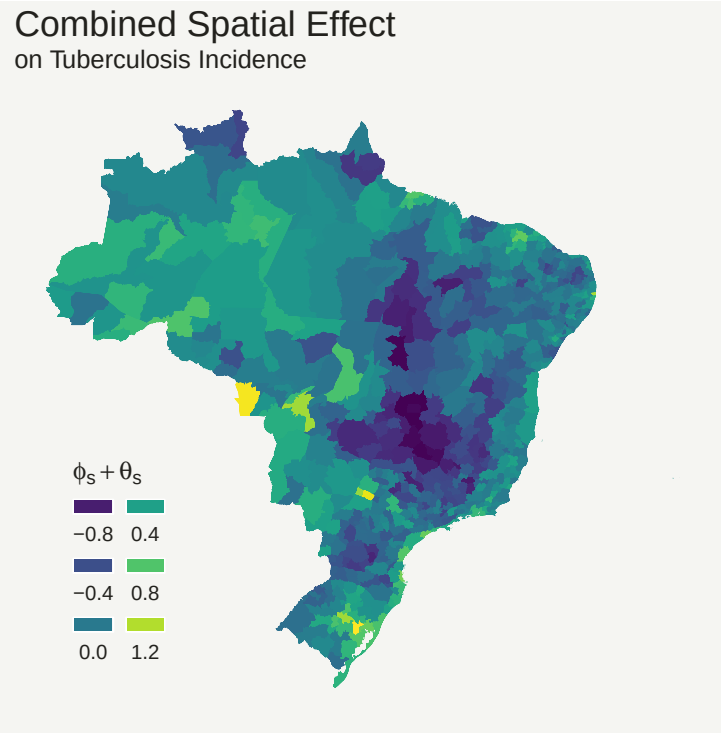

Figure 2.11: Combination of structured spatial effect ϕ_s and unstructured effect θ_s.

Figure 2.12 shows a clear, monotonically increasing (estimated) relationship between treatment timeliness and the probability of reporting $\pi_{t,s}$. The 95% CrI does not incorporate a horizontal line, which would imply no relationship. Overall, micro-regions with very low timeliness ($< 10\%$) have approximately two-thirds the reporting probability of ones with very high timeliness ($> 90\%$), indicating a clear disparity in the performance of the surveillance programs.

Finally, Figure 2.13 shows, for each year, the total observed TB count, alongside the 5%, 50% and 95% quantiles of the predicted true total number of unreported cases. The plot suggests that potentially tens of thousand of cases went unreported each year. Combined with the results seen in Figure 2.12, this presents a strong case for providing additional resources to the surveillance programs in those micro-regions with lower values of treatment timeliness. The R code and data needed to reproduce these results are provided in the Supplementary Material.

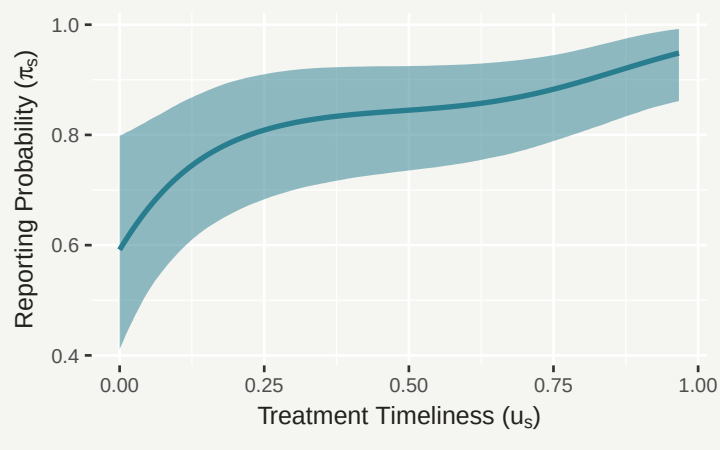

Figure 2.12: Posterior mean predicted effect of treatment timeliness on the reporting probability of TB, with associated 95% CrI.

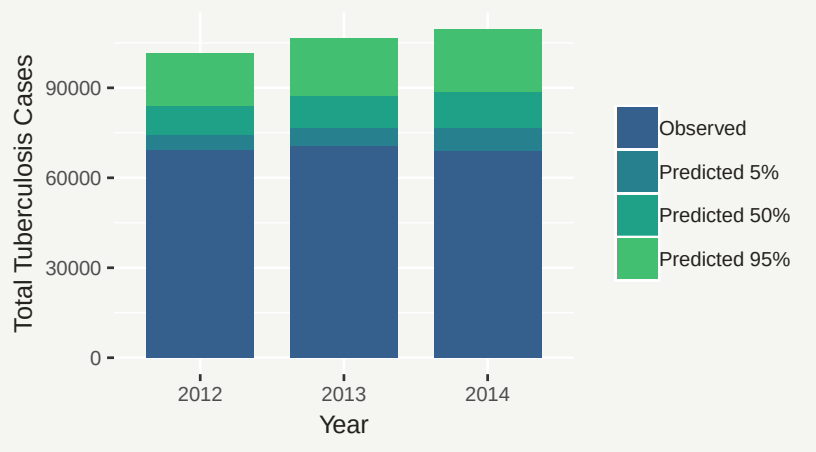

Figure 2.13: Bar plot showing, for each year, the recorded total number of TB cases in Brazil, as well as the 5%, 50% and 95% quantiles of the predicted true total number of TB cases.

2.4 Application to UK Tornado Data

Tornados pose a serious risk to society through damage they can cause to property and infrastructure, such as power stations. Tornado data in the UK are collected by the Tornado and Storm Research Organisation (TORRO). However, due to the localised nature of tornados, the reporting process relies on direct observation by the local population and/or media, or indirect observation through the examination of resulting damage. For this reason, it is reasonable to suppose that the probability of each tornado being reported may be lower in sparsely populated areas than in areas with a larger number of potential observers. If this is the case, then this could lead to the under-estimation of the risk posed by tornados in areas with low population densities. To better understand this risk, we apply the hierarchical model presented in Section 2.2.2 and in Stoner et al. (2019a) to reported tornado counts, aggregated over the period 1950 to 2010 and arranged into a grid of 25×25km cells across Great Britain and some surrounding islands.

2.4.1 Methodology

For grid cell $s \in \{1, ..., 468\}$, let z_s denote the total number of tornados reported over the period 1950 to 2010 and let y_s denote the true number of tornados which occurred

in the same period. In the model for tuberculosis counts presented in Section 2.3, an independent identically distributed Normal random effect was included in the mean of the Poisson model for the true number of tuberculosis cases. This effect was included to afford the Poisson model additional variation which may be caused by the effect of potential covariates which weren't included in the model. An alternative approach which has a near equivalent interpretation is to replace the Poisson model with a Negative Binomial model, of which the Poisson distribution is a limiting case, which includes a second parameter θ to allow for additional variance.

$$\begin{aligned} z_s \mid y_s &\sim \text{Binomial}(\pi_s, y_s) && (2.34)\\ \log\left(\frac{\pi_s}{1-\pi_s}\right) &= \beta_0 + g(d_s) && (2.35)\\ y_s &\sim \text{Negative-Binomial}(\lambda_s, \theta) && (2.36)\\ \log(\lambda_s) &= \log(A_s) + \alpha_0 + f_1(s) + f_2(\nu_s) && (2.37) \end{aligned}$$

To capture the relationship between population density and the reporting rate, we included a cubic polynomial $g(d_s)$ of the logarithm of population density d_s (from the 2010 national census) in the model for the Binomial reporting probability π_s. It is believed that terrain has a substantial impact on tornado occurrence (Elsner et al., 2016). To capture this, we include in the tornado rate model a one-dimension thin-plate spline $f_2(\nu_s)$ of the mean altitude ν_s within each grid cell s. In the data the number of recorded tornados is broadly higher in the south-east of Great Britain than in the north-west (Kirk, 2014). It is possible that at least some of this discrepancy is caused by under-reporting, as population density is higher on average in the south-east. The difference may also be due to different terrain. It is possible, however, that even after accounting for these two effects there may be some remaining spatial variation in tornado incidence, for example due to climatological variation. To potentially capture this, we include a two-dimensional thin plate spline $f_1(s)$ of spatial location s, with the two dimensions being the northing and easting coordinates of the grid cell, in the model for the expected tornado incidence. Finally, the logarithm of the total area A_s within grid cell s made up of land (as opposed to water) was included as an offset in the model for tornado incidence, so that we are effectively modelling the tornado rate per squared kilometre of land.

As discussed in Sections 2.2-2.3, as we do not have any counts which we know do not suffer from under-reporting, we need to provide some form of prior information for the reporting rate to achieve identifiability between the reporting rate and the incidence rate. In this case, our prior belief is that the reporting rate is very likely close to 100% in areas with the highest population densities. To incorporate this belief in the model, we applied a linear transformation to the population density covariate d_s so that the cubic polynomial $g(d_s)$ is zero at the maximum value of population density in the data. This means that the intercept β_0 has the interpretation of the reporting probability (at the logistic level) in the most densely populated

areas. If we wished to fix the reporting rate to be equal to 100% at the highest population density value, we could alternatively use a log link for π_s, as described in Section 2.2.2 with $\beta_0 = 0$. To instead incorporate at least some degree of uncertainty in the maximum reporting rate, we retained the logistic link and placed a strong Normal$(5, 1)$ prior on β_0. This equates to a prior distribution for π_s where, in the most densely populated areas, the most likely value is just over 99%, with a very low probability ($< 2\%$) of the maximum reporting probability being less than 95%.

All code was written and executed using R (R Core Team, 2018) and the model was implemented using NIMBLE (de Valpine et al., 2017). The one-dimensional and two-dimensional thin plate splines were set up using the `jagam` function (Wood, 2016). For MCMC efficiency, the model was implemented using the marginal model for the recorded count (z_s) presented in (2.9). To improve sampling efficiency, an automated factor slice sampler (Tibbits et al., 2014) was used to jointly sample α_0, β_0 and the population density polynomial coefficients. Four MCMC chains were run from different randomly generated initial values and with different random number generator seeds for a total of 20k iterations, discarding 10k as burn-in. Convergence of the MCMC chains was assessed by computing the Multivariate Potential Scale Reduction factor for the intercepts, polynomial coefficients, spline coefficients and dispersion parameter θ. A value of 1.03 was obtained, suggesting the chains had converged.

2.4.2 Results

For reasons of data confidentiality, we do not present any figures from which the original counts are recoverable. Additionally, comments from TORRO did not influence the model or results but did provide some useful context for the geographical variation seen in the data. Figure 2.14 shows the posterior mean effect $f_1(s)$ of climate on tornado incidence. A strong gradient can be seen from south-east to north-west, with the mean tornado rate over 4 times higher in south-east England compared to Scotland.

Figure 2.15 shows the posterior mean effect $f_2(\nu_s)$ of mean altitude on tornado incidence, with associated 95% intervals. Mean tornado incidence decreases substantially as altitude increases, with over double the incidence rate at sea level compared to high altitude (over 300m) areas. The tightness of the 95% intervals suggests this relationship is very strong.

An even more interesting picture can be seen when the location ($f_1(s)$) and altitude ($f_2(\nu_s)$) effects on the mean tornado rate are combined, as shown in Figure 2.16. Looking at the posterior means of the combined effects, the tornado rate is potentially 20 times higher in low-lying areas in south-east England compared to higher areas in Scotland.

Figure 2.17 shows the posterior mean effect of population density on the Binomial

reporting probability π_s. We can see that the reporting rate is very likely to be less than 25% in the least densely populated areas.

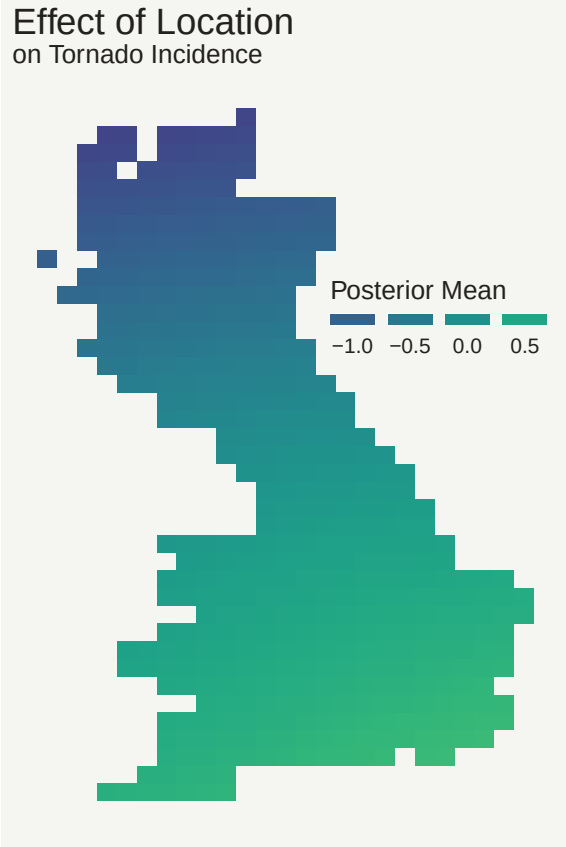

Figure 2.14: Posterior means of the two-dimensional (northing and easting) thin plate spline effect $f_1(s)$ of spatial location s, designed to capture the effect of geographic location on tornado incidence.

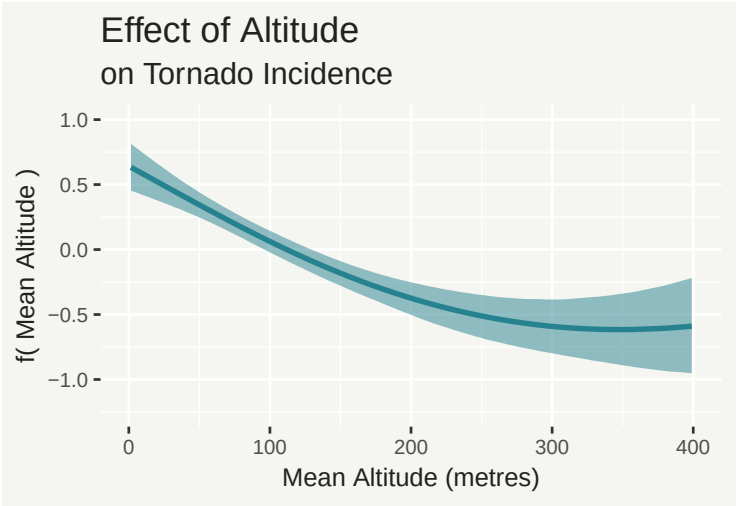

Figure 2.15: Posterior means of the thin-plate spline effect $f_2(\nu_s)$ of mean altitude on tornado incidence, with associated 95% credible intervals.

2.4.3 Conclusion

In a way, this is an ideal application of the hierarchical framework we have presented, because it is relatively clear that population density belongs in the model for the reporting probability and not the model for tornado incidence.

From the model we have discovered that in some parts of the UK as much as 75% of all tornados could have gone unreported in the period 1950-2010. This implies

that ignoring the under-reporting could result in a severe under-estimation of the risk posed by tornados in areas of low population density, which has implications for policy such as the planning of new nuclear power stations, which are often positioned in remote areas. Moreover, taking into account the under-reporting has allowed us to learn more about the effect of both altitude and location on tornado incidence. In particular, if we had not taken into account the effect of population density on the reporting rate, our spatial inference would likely have been considerably biased. This is because in the UK population density tends to be much higher on average in the south-east of England than in the rest of the country.

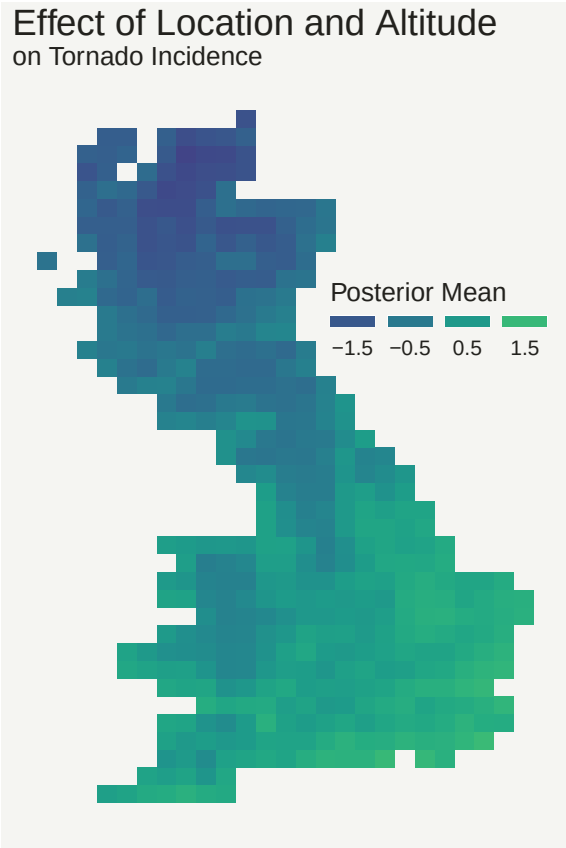

Figure 2.16: Posterior means of the combination of the two-dimensional thin-plate spline effect $f_1(s)$ of location and the thin-plate spline effect $f_2(\nu_s)$ of mean altitude on tornado incidence.

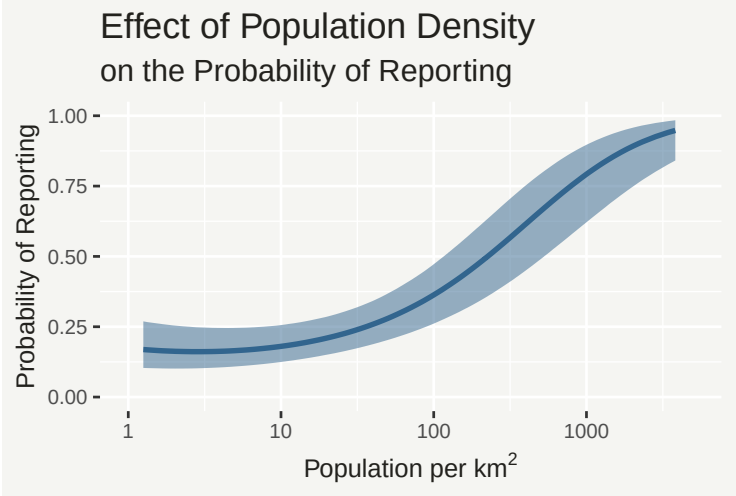

Figure 2.17: Posterior means of the effect of population density on the reporting probability π_s, with associated 95% credible intervals.

2.5 Application to Historic Volcano Data

Note that a concise version of this section has been published in Stoner (2018). Historical volcano data can suffer from under-recording of eruption occurrence, which can vary with time and magnitude. We once again employ the hierarchical framework for correcting under-reporting, to model simultaneously the true eruption rate and the under-recording mechanism, in order to obtain a more reliable inference on the relationship between eruption magnitude and frequency.

2.5.1 Introduction

The LaMEVE dataset is a record of historic eruptions, with each entry including an estimated eruption year and an estimate of the magnitude of the eruption. This is defined by:

$$\text{magnitude} = \log_{10}(\text{erupted mass in kg}) - 7$$

Unfortunately, the recording of volcano eruptions is not complete; some entries rely on historical records, while many rely on geological analyses, where the likelihood of an eruption leaving a discoverable trace depends on the location, time and magnitude of the eruption (Rougier et al., 2018).

This means that any inference on the temporal profile of eruption rates which assumes complete recording is likely biased. It is therefore desirable to quantify the under-recording, such that the frequency of eruptions, and its relationship with magnitude, can be more reliably investigated.

2.5.2 Methodology

The dataset is a list of eruptions, with columns to indicate the year in which they occurred and an estimate of magnitude. In order to use the hierarchical framework for correcting under-reporting presented in Section 2.2.2 and in Stoner et al. (2019a), the data must be aggregated into counts of eruptions over a chosen time interval and to achieve this the data were aggregated over 1000 intervals of 100 years.

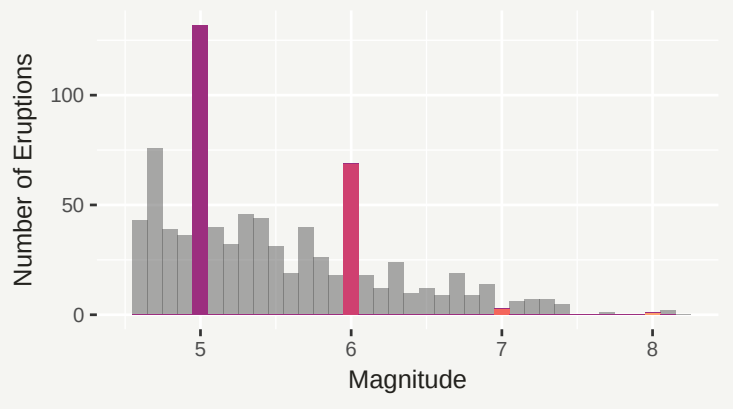

Figure 2.18: Histogram of recorded eruption magnitudes within the last 1000 centuries.

As discussed in Rougier et al. (2018), a large portion of eruption magnitude estimates have been rounded to the nearest integer. Figure 2.18 shows a histogram of recorded mangnitude estimates within the last 1000 centuries. It can be clearly seen that the number of magnitude estimates recorded either as exactly 5 or exactly 6 rise well above the distribution of the other values. If ignored then this could introduce issues if eruption magnitude is treated as a continuous variable, such as potential overestimation of the rate of near-integer value magnitudes and underestimation elsewhere. To avoid this problem, we follow Rougier et al. (2018) by classifying the data into four bins based on magnitude: Low [4.5,5.5), Medium [5.5,6.5), High [6.5,7.5) and Very High [7.5,8.5). The resulting set of eruption counts can be seen in Figure 2.19. Ignoring the under-recording, the data appear to suggest the rate of eruptions has been dramatically increasing in recent centuries.

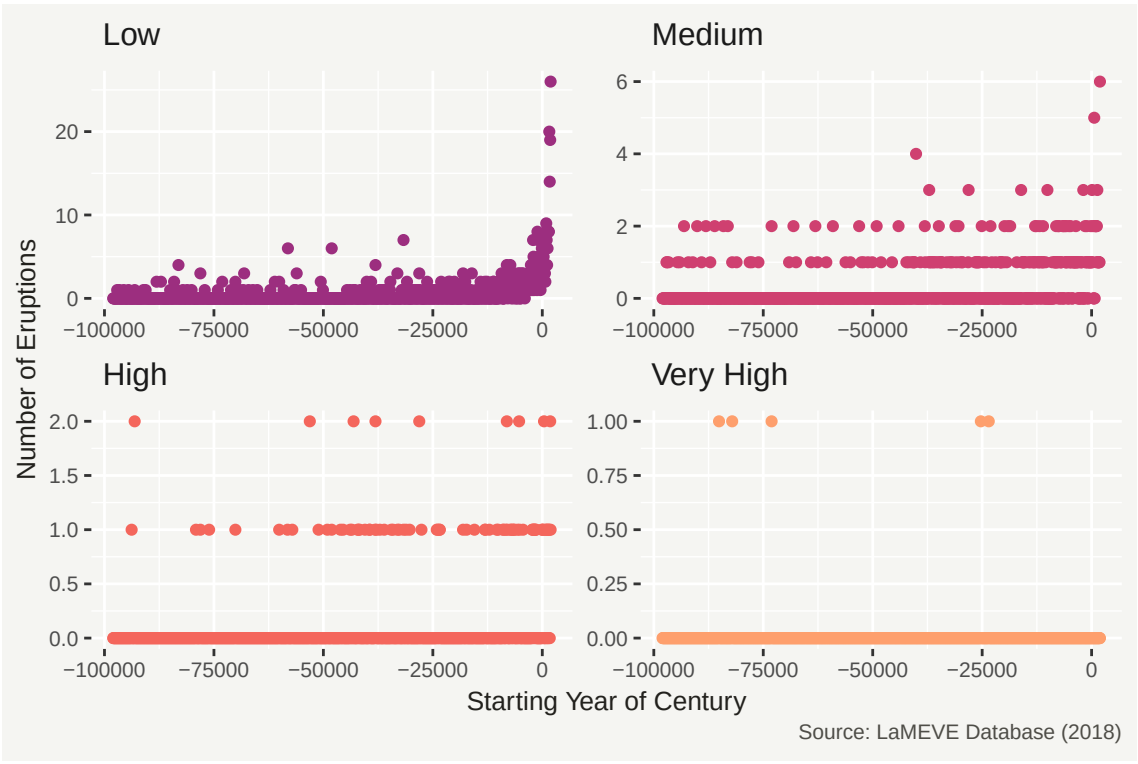

Figure 2.19: Total eruption counts for each of the last 1000 centuries, by magnitude.

For an eruption in century $t \in T = \{1, ..., 1000\}$, where $t = 0$ represents the 21st century and going backwards in time so that $t = 1$ represents the 20th century, and of magnitude $m \in M = \{\text{Low, Medium, High, Very High}\}$ the model for the

recorded number of eruptions $z_{t,m}$ is given by:

$$
\begin{aligned}
z_{t,m} \mid y_{t,m} &\sim \text{Binomial}(\pi_{t,m}, y_{t,m}) && (2.38)\\
\log\left(\frac{\pi_{t,m}}{1-\pi_{t,m}}\right) &= \beta_{0,m} + \sum_{k=1}^{3} \beta_{k,m} w_{t,m}^{k} && (2.39)\\
y_{t,m} &\sim \text{Poisson}(\lambda_m) && (2.40)\\
\mu_m &= \alpha_0 + \alpha_1 x_m && (2.41)\\
\log(\lambda_m) &\sim \text{Normal}(\mu_m, \sigma^2) && (2.42)
\end{aligned}
$$

Here $w_{t,m}$ is the transformed century t ($w_{t,m} = \log(t+1)$ for $m = \text{Low}$, $w_{t,m} = t/1000$ otherwise), and x_m is defined by the midpoint of the magnitude bin, minus the mean of the midpoints. The change in the recording probability $\pi_{t,m}$ is characterised by a different cubic polynomial for each magnitude bin in (2.39). A log-linear relationship between the eruption rate and magnitude is introduced in (2.41), such that information is pooled from the different bins into parameters α_0 and α_1. This is intended to aid in estimating the rate of Very High eruptions, of which there are very few observations. Additional flexibility is introduced by allowing the eruption rate for each bin to deviate from this line according to a Normal distribution, to allow for potential biases in the estimation of eruption magnitude.

We specified relatively non-informative Normal$(0, 10^2)$ prior distributions for the rate intercept and slope parameters (α_0 and α_1), and Normal$(0, \sigma^2 = 1000)$ priors for the polynomial coefficients $\beta_{k,m}$ ($k \in \{1, 2, 3\}$). For the variance parameter σ we specified a Gamma$(2, 1)$ prior distribution, reflecting our belief that, while very high variances are unlikely, the rates for magnitudinal bins are likely to have at least a modest amount of variance about the mean line $\alpha_0 + \alpha_1 x_m$, due to issues such as rounding. As there are no available observations of the true eruption counts $y_{t,m}$, the model is non-identifiable between a high eruption rate λ_m and a low reporting probability $\pi_{t,m}$ or vice-versa. In the tornado application, we were able to rectify this by using an informative prior that the reporting probability is very close to 100% in the most densely populated areas. Similarly, we might believe that the reporting rate of all eruptions is close to 100% in the 21st century ($t = 0$). Note that the right hand side of the model for $\pi_{t,m}$ given in (2.39) reduces to $\beta_{0,m}$ when $t = 0$ (as $w_{0,m} = 0$ by definition for all magnitudes). This means that $\beta_{0,m}$ can be interpreted as the 21st century reporting rate. To incorporate our belief that this rate is likely to be close to 100%, we placed an informative Normal$(4, \sigma^2 = 1/2)$ prior on the $\beta_{0,m}$ for Low, Medium and High magnitudes. This corresponds to a prior for the 21st century reporting rate where the most likely value is approximately 98%, with a very low (<1%) probability that it is less than 90%. However, we believe that the number of non-zero observations (only 5) for the Very High bin is too low to provide any meaningful inference for the change in recording probability over time, so we fixed the recording probability at 1 for this bin.

All code was written and executed using R (R Core Team, 2018) and the model was implemented using NIMBLE (de Valpine et al., 2017). As in the tornado model, for MCMC efficiency the model was implemented using the marginal model for the recorded count ($z_{t,m}$) described in (2.9). To improve sampling efficiency, an automated factor slice sampler (Tibbits et al., 2014) was used for each $m \in M$ to jointly sample the reporting probability intercept and polynomial coefficients $\boldsymbol{\beta}_m$ and the log-rate γ_m. Four MCMC chains were run from different randomly generated initial values and with different random number generator seeds for a total of 20k iterations, discarding 10k as burn-in. Convergence of the MCMC chains was assessed by computing the Multivariate Potential Scale Reduction factor for the intercepts α_0 and $\boldsymbol{\beta}_0$, the reporting probability polynomial coefficients, the slope parameter of the rate model α_1 and the variance parameter for the rate model σ. A value of 1.01 was obtained, suggesting the chains had converged.

2.5.3 Results

The estimated change in the recording probabilities, for the first three magnitudinal bins, can be seen in Figure 2.20. All three curves show near monotonic decreasing trends going backwards in time, with a pattern of increasing recording probability for higher magnitudes generally holding.

Figure 2.21 shows the posterior mean estimates of the eruption rate for each magnitudinal bin, with associated 95% credible intervals. The solid black line represents the median predicted relationship between magnitude and rate, as defined in (2.41).

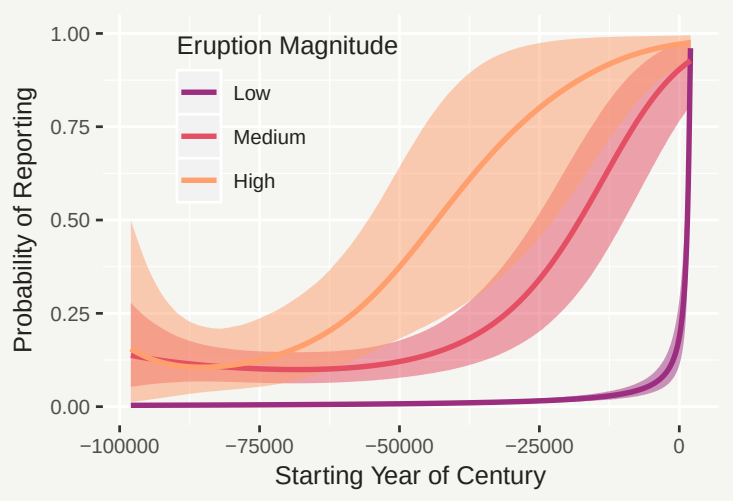

Figure 2.20: Posterior mean estimates of the effect of time on the probability of recording a volcano eruption, for the first three magnitude bins, with associated 95% credible intervals.

Finally, the return period R_m, the expected number of years in between occurrences, for an eruption in a given magnitude classification m can be calculated as:

$$R_m = \frac{1}{\lambda_m} \tag{2.43}$$

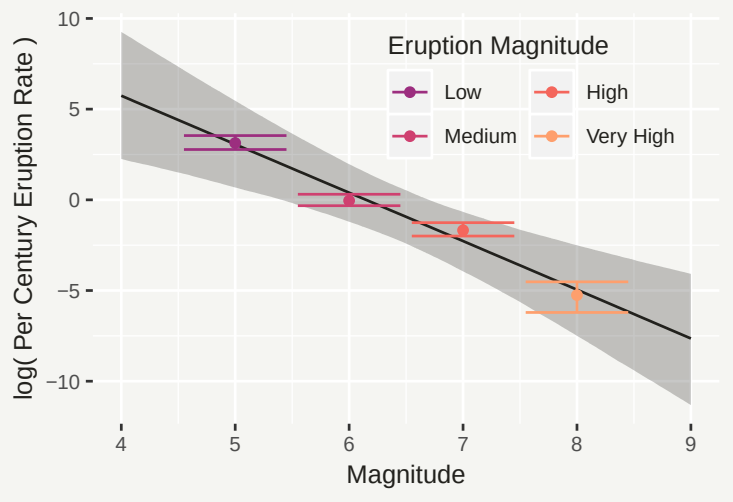

Figure 2.21: Estimated eruption rate (on the log-scale) for the different magnitudinal bins, with associated 95% credible intervals. The solid black line shows the posterior median estimate of $\alpha_0 + \alpha_1 x_m$, with associated 95% credible interval.

In Table 2.1, it can be noted that the return period estimate for a Very High eruption is similar to the estimate in Rougier et al. (2018) for the return period of a volcano exceeding magnitude 8, which has a median of 17000 years, with associated 95% credible interval (5200,48000).

Magnitude	Lower 95%	Median	Upper 95%
Low	2.9	4.4	6.2
Medium	74	100	140
High	350	530	740
Very High	9200	19000	50000

Table 2.1: Return period (years) approximate predictive quantiles for an eruption in each of the four magnitudinal bins.

2.5.4 Conclusion

In this section the challenges posed by under-recording were explored in the context of historic volcano eruptions. A dataset of volcanic eruptions was aggregated both by century and into four bins, based on eruption magnitude, which resulted in a dataset of counts. To this dataset we applied the general framework for correcting under-reporting presented in Stoner et al. (2019a). Non-identifiability was resolved using informative priors for the 21st century reporting rates for each magnitude, representing the belief that they are near 100%.

By accounting for the relationship between eruption magnitude and time and the under-recording mechanism, a more reliable inference for the relationship between eruption magnitude and rate of occurrence was made possible. In particular, our results suggest the reporting rate increases with magnitude. This implies that ignoring the under-reporting mechanism could have led us to infer that, relative

to higher magnitude eruptions, low magnitude eruptions occur less frequently than they actually do.

This inference relies on the assumption that the rate of eruptions was constant in time over the period analysed, and the results could also be sensitive to the way in which the eruptions were aggregated both into centuries and into four magnitudinal bins. Both of these issues are worthy of future investigation, though notably our results were similar to those in Rougier et al. (2018), a study of the same data set with a substantially different approach.

2.6 Further Simulation Experiments

2.6.1 Informative prior versus completely observed counts

In Section 2.2 we discussed the need to supplement the lack of information in the data, in order to distinguish between the under-reporting rate and incidence rate. This is done by either providing an informative prior distribution for β_0, the mean reporting rate at the logistic scale, or by utilising some completely reported counts, or both. In this experiment we investigate the effect of varying the strength of the informative prior and the number of completely observed counts, on predictive uncertainty.

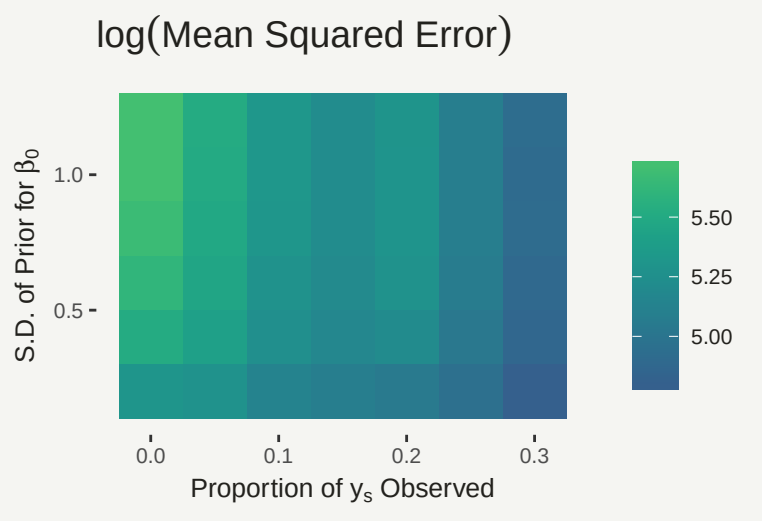

Figure 2.22: Mean values of the posterior predictive log-mean squared errors for each modelling scenario.

The model was applied to simulated data, as in Section 2.3.3, using different values for the prior standard deviation, to reflect varying levels of prior certainty about the reporting rate, and including completely reported counts for varying proportions of the data. Predictive uncertainty was quantified using the logarithm of the mean squared error of y_s, computed for each posterior sample, which we summarise using the mean. Figure 2.22 shows how this uncertainty varies with prior variability in β_0 and the number of completely reported counts. The left-most column shows that predictive uncertainty decreases with increasing prior precision when there are no

completely reported counts. In this case, practitioners must trade-off predictive uncertainty with the risk of systematic bias posed by specifying an overly strong prior away from the true value, seen in Section 2.3.3. While predictive uncertainty does decrease with increasing prior strength, we can also see that it decreases more substantially by increasing the proportion of counts which are known to be completely reported. This implies that the use of completely observed counts is worthwhile, if possible.

2.6.2 Strength of under-reporting covariate

In Section 2.3.3, we varied the strength of relationship between the under-reporting covariate and the true under-reporting covariate. Figure 2.23 shows the relationship between the different "proxy" covariates and the reporting probability π_s. This section presents the effect of using these proxies instead of the true under-reporting covariate w_s.

The three plots in Figure 2.24 summarise the effect that varying the strength of this covariate has on the performance of the model, using locally weighted scatterplot smoothing (LOESS). The left plot shows the 95% PI coverage. As discussed in Section 2.3.3, coverage should not decrease with covariate strength, and indeed there is very little evidence of any change. The central plot shows the mean error of $\log(\lambda_s)$. Again, the plot shows little evidence that this changes with covariate strength, which is reassuring as it suggests that using a weaker covariate does not necessarily introduce any systematic bias. Finally, the right plot shows a substantial effect of covariate strength on the predictive accuracy of $\log(\lambda_s)$, with stronger covariates translating to higher predictive accuracy, which is expected.

This experiment suggests that gains in predictive accuracy can be achieved by using covariates that are only proxies of the under-reporting process, compared to not including them, without necessarily introducing bias. However, this relies on those covariates being correctly identified as being related to the under-reporting mechanism. The following section illustrates the risks associated with this classification.

2.6.3 Classification of covariates

In the application to TB data, the classification of covariates into those that relate to the under-reporting mechanism and those related to the true count generating process was relatively straightforward. In general, this can be more challenging and in this section we present the effects of incorrectly classifying covariates.

The experiment begins by using simulated data from the model in Section 2.3.3, with the exception of an additional unstructured random effect in the model for λ. The prior distributions are the same, with a $\mathrm{N}(0, 0.6^2)$ prior on β_0. In the first

instance, the model is correctly informed that covariate x_s belongs in the model for λ_s and w_s belongs in the model for π_s. In the second instance, these are swapped. For comparison, the model is also applied with no covariates included.

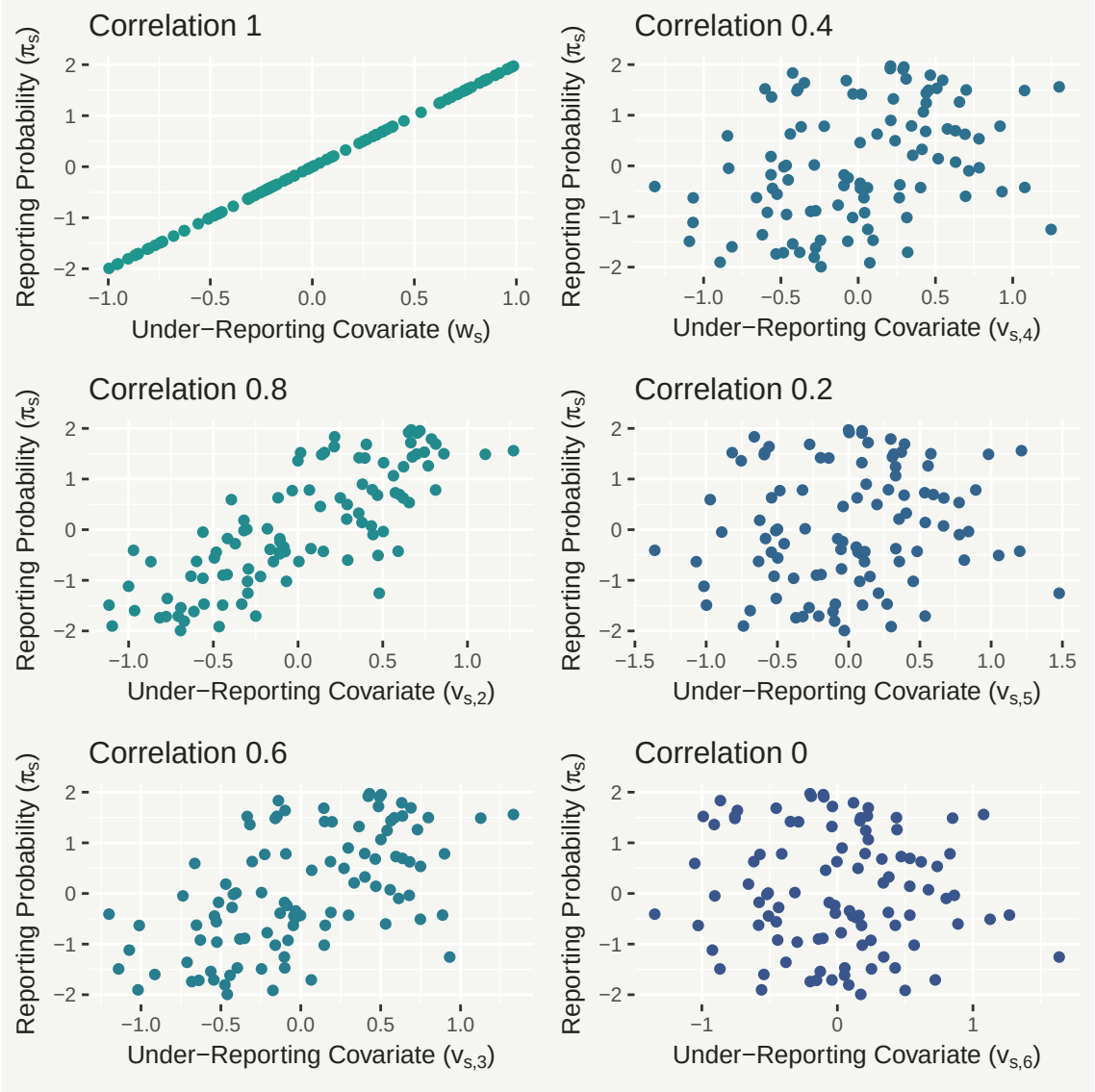

Figure 2.23: Scatter plots comparing covariates $v_{s,2}, ..., v_{s,6}$ to the reporting probability π_s.

Figure 2.25 shows scatter plots for each case, comparing median predicted values for y_s to their corresponding true values. The left plot shows that when the covariates are correctly classified, the model is able to detect the unobserved y_s values very well. When the covariates are incorrectly classified (right), the model performs very poorly. In fact, in this case the model performs even worse than a model where no covariates are included and the only random effects are relied upon to improve predictions (centre).

This experiment highlights the sensitivity of the framework to the classification of covariates, which represents an informative choice. In our view, if there is substantial doubt about whether a covariate likely relates to the under-reporting mechanism or

to the true count process, it may be wiser to not include it in the model, which in this experiment results in better predictive performance.

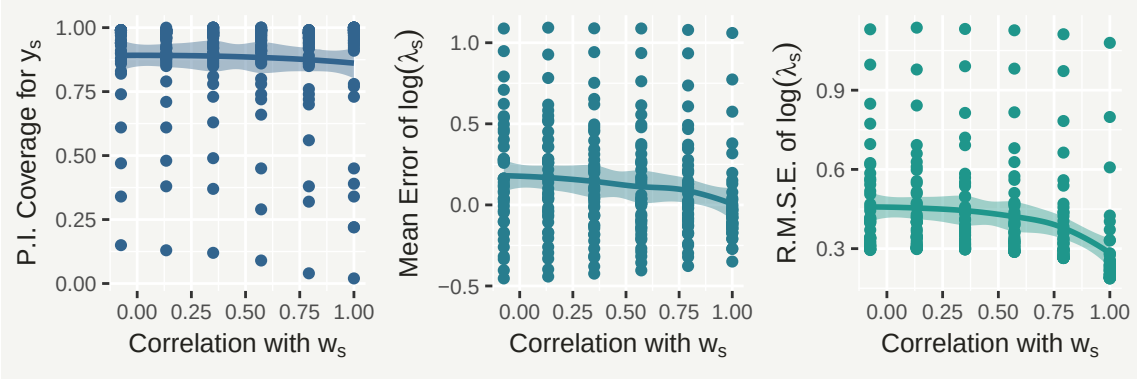

Figure 2.24: Scatter plots comparing the correlation of the under-reporting covariate used, from the set $v_{s,1}, ..., v_{s,6}$, to 95% PI coverage for the true counts y_s (left), the mean error of $\log(\lambda_s)$ (centre) and the square root of the mean squared error of $\log(\lambda_s)$ (right).

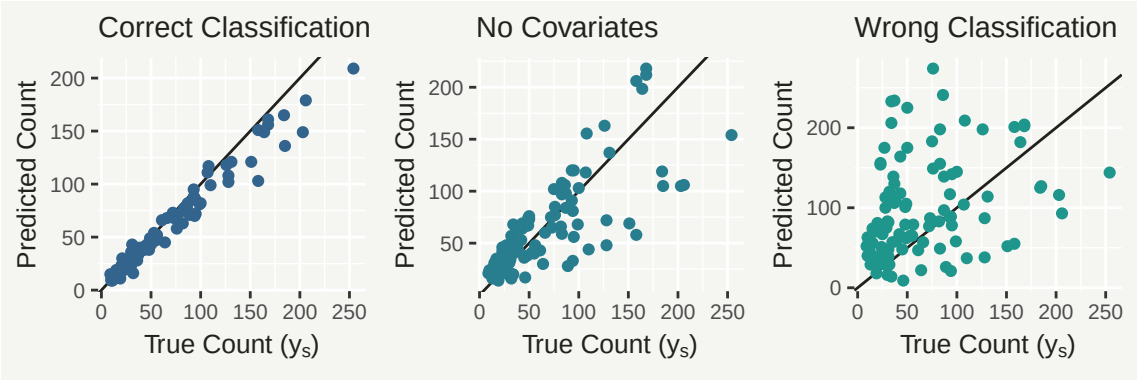

Figure 2.25: Scatter plots comparing the true simulated counts y_s to the median predicted counts from the model where the covariates are classified correctly (left) and incorrectly (right), and the model where the covariates are not included (centre).

2.6.4 Other simulation experiments

Although the experiments we present give a useful insight into the sensitivity of our proposed framework under varying modelling scenarios, they were limited in scope owing to issues of practicality.

Specifically, in our experiments we covered three dimensions: the mean of the prior for β_0, the standard deviation of the prior for β_0 and the strength of correlation between the available under-reporting covariate and the true under-reporting covariate). This resulted in 252 models to run, taking around 8 hours on a high-end desktop computer in 2018. In addition to these three dimensions, we were also interested in investigating, for example, the effect of the strength of correlation between

the under-reporting covariate w_s and the process covariate x_s, though this would have increased the time and resources required for computation.

All of these experiments would still be based on just one simulated dataset, from one set of parameter values. A truly comprehensive simulation study where these are varied is therefore computationally impractical at the time of submission, but may be possible in the future.

2.7 Discussion

A flexible modelling framework for analysing potentially under-reported count data was presented. This approach can accommodate a situation where all the data are potentially under-reported, by using informative priors on model parameters which are easily interpretable. It also readily allows for random effects for both the count process and the under-reporting process, something which simulation experiments revealed alleviates the use of proxy covariates to determine under-reporting rates. It was applied primarily to correcting under-reporting in TB incidence in Brazil using well-established MCMC software, incorporating a spatially structured model which highlights its flexibility. Simulation experiments were conducted to investigate prior sensitivity and to provide a guide for choosing a prior distribution for the mean reporting rate.

Naturally, care should be taken. Indeed, it is likely that a different prior distribution for β_0 in the TB application might result in different inference on the under-reporting rate, and consequently the corrected counts. The simulation experiments indicated that if the specified prior information on the overall under-reporting rate turns out to be wildly different from the truth, then the corrected counts will also likely be inaccurate. Therefore particular attention should be paid to the elicitation of this prior information, such that the prior uncertainty is fully quantified and reflected in predictive inference. Further simulation experiments also highlighted the risk posed by incorrectly classifying covariates as either belonging in the under-reporting mechanism or the model of the true count. In many cases strong prior information about this classification may be available, so we suggest future research is directed at combining prior uncertainty with methods such as Bayesian model averaging. This could more rigorously quantify the uncertainty associated with this classification and its effect on the predictive inference for the corrected counts.

The subjective nature of the solution to completely under-reported data is not unique; in Bailey et al. (2005) for example, a different choice of threshold for the variable used to identify under-reported counts could have lead to different predictions. Only the usage of a validation study (e.g. Stamey et al. (2006)) could be considered a less subjective approach depending on the quality, quantity and experimental design of collecting the validation data. In many cases however, the

elicitation of an informative prior distribution for one parameter is simply a more feasible solution. In the application to TB, an existing estimate from the WHO of the overall reporting rate in Brazil was available, from which a prior distribution was derived.

The framework investigated here has two key advantages over the approaches based on censored likelihood discussed in Section 2.2.1. Firstly, modelling the severity of under-reporting, through the reporting probability, presents the opportunity to reduce under-reporting in the future, by informing decision-making about where additional resources for surveillance programmes would be most effective. Secondly, by modelling the under-reported counts, a more complete predictive inference on corrected counts is made available, informed by the reporting probability, the rate of the count-generating process and the recorded count. The results in Section 2.3, for instance, provide predictions of the under-reporting rate at a micro-regional level, meaning that resources could be intelligently applied to the worst-performing areas.

Chapter 3

Household Air Pollution

This chapter is largely based on Stoner et al. (2019b) which was under review at the time of thesis submission.

Globally, an estimated 3.8 million deaths per year can be attributed to household air pollution from the combustion of solid fuels and kerosene for cooking. Information on the proportion of people relying primarily on different polluting fuels for cooking is available in the form of nationally-representative household surveys. However, the absence of a modelling framework for comprehensively estimating the use of individual fuels inhibits fuel-specific policy interventions. To address this, we develop a multivariate hierarchical model (known as the Global Household Energy Model, or GHEM) for data from the World Health Organization Household Energy Database, spanning the period 1990-2016.

Based on Generalized-Dirichlet-Multinomial distributions, the model jointly estimates trends in the use of eight individual fuels, whilst addressing a number of challenges involved in modelling the data. These include: missing values arising from incomplete surveys; missing values in the number of survey respondents; and sampling bias in the proportion of urban and rural respondents. The model also includes regional structures to improve prediction in countries with limited data. We assess model fit using within-sample predictive analysis and conduct an out-of-sample prediction experiment to evaluate the model's forecasting performance. Overall, this work substantially contributes to expanding the evidence base for household air pollution, which is crucial for developing policy and planning interventions.

3.1 Introduction

In 2018, the World Health Organization (WHO) estimated that about 7 million deaths globally could be attributed to air pollution. This total comprises deaths associated with the joint effects of ambient (outdoor) and household exposure to air pollution. For household exposures, the proportion of households in a country relying mainly on various polluting fuels and technologies for cooking is used

as an indicator of population exposure to household air pollution. In accordance with WHO guidelines for indoor air quality (household fuel combustion), households mainly cooking with coal, wood, charcoal, dung, crop residues or kerosene are considered exposed (World Health Organization, 2014). Globally, it is estimated that in 2016, 41% of the world's population were exposed to household air pollution resulting from cooking with polluting fuels and technologies (World Health Organization, 2018c). The use of polluting fuels and technologies for cooking is primarily a problem in low and middle income countries where little over half (52%) of people have access to clean fuels, compared with 99% in high income countries.

Although in South East Asia the proportion of the population relying on clean fuels and technologies has doubled over the last two decades, progress in the African region has been much slower, with less than a 4% increase in the population using clean fuels and technologies for cooking (World Health Organization, 2018c). Despite the apparent increase in the percentage of the population using clean fuels and technologies for cooking, the absolute number of people without access to clean fuels and technologies has stayed fairly constant. Global figures have remained unchanged since 2000, with currently over 3 billion people still relying on polluting fuels and technologies for cooking. To make further progress, through interventions such as encouraging households to adopt cleaner fuels like liquid petroleum gas (LPG) or promoting the use of technologies which make cooking with polluting fuels safer, it is essential to understand current and past temporal trends in fuel use.

The proportion of populations with primary reliance on clean fuels and technology serves as a key indicator (7.1.2) for monitoring progress towards the Sustainable Development Goal (SDG) 7.1 '...to ensure universal access to affordable, reliable and modern energy services'. It also forms an important part of indicator 3.9.1, the mortality rate attributed to the joint effects of ambient and household air pollution, which monitors progress towards SDG 3.9, '... to substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination'. Since the year 2000, all regions have seen progress in access to clean household energy but at varying rates.

Here, we propose a model that estimates trends in the proportion of people relying on individual fuels for cooking in each country, together with associated measures of uncertainty. Relationships between the use of different fuel types are modelled together with the variability associated with survey sampling, which may vary by country. Where data is not available within a country, or is insufficient to produce accurate estimates, the model structure derives information from regional trends. The model allows for different fuel usage in urban and rural areas and is able to produce predictions (with associated uncertainty) of future use of different fuel types, providing policy makers with a baseline against which they can evaluate the effectiveness of future interventions.

The remainder of the chapter is organized as follows: Section 3.2 discusses the available data and previous approaches to modelling household fuel use; Section 3.3 provides details of the proposed modelling approach, including the implementation of the model using Markov chain Monte Carlo (MCMC); and Section 3.4 presents posterior predictive model checking and a future forecasting experiment. Finally, Section 3.5 provides an overall summary and a concluding discussion of the model's impact.

3.2 Background

Information on the types of technologies and fuels used by households for cooking is regularly collected in nationally-representative household surveys or censuses. These data are compiled in the WHO Household Energy Database (World Health Organization, 2018a) which, as of late 2018, contains over 1100 surveys, with data from 157 countries for the years 1990 to 2016. Survey data are used to calculate the annual percentage of households in each country which use either solid fuels (charcoal, coal, crop waste, dung, rubber and wood) as their primary cooking fuel, liquid fuels (kerosene) or others including gaseous fuels or electricity.

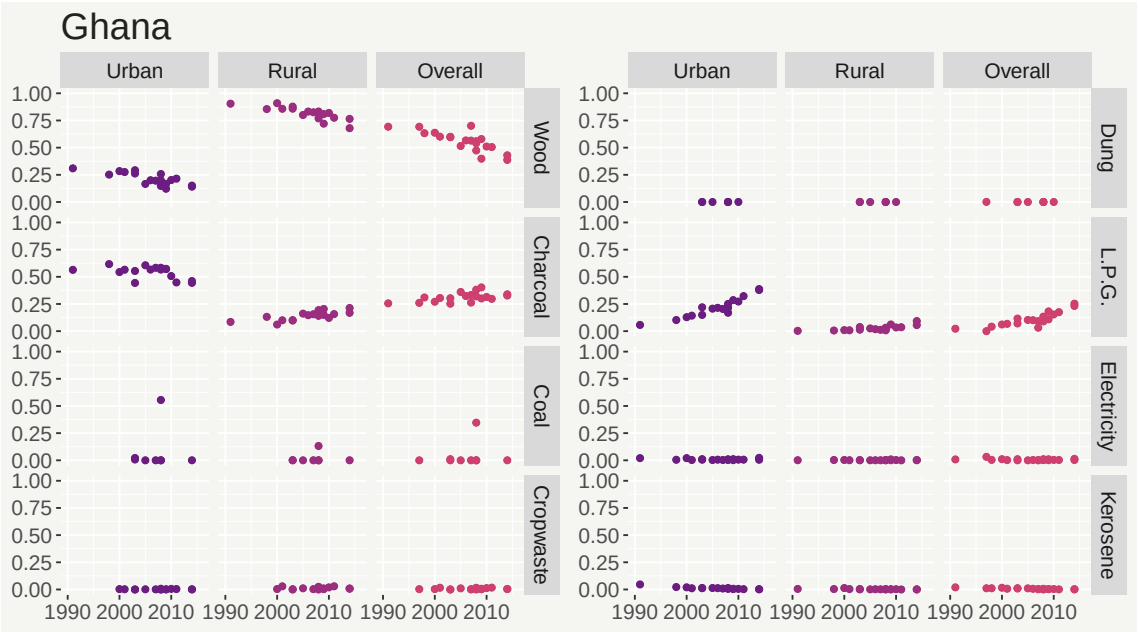

Figure 3.1: Time series of fuel use from surveys of Ghana, in a 2018 snapshot of the WHO Household Energy Database.

Figures 3.1 and 3.2 show example time series of survey data for Ghana and Mongolia, respectively. We can see that for Ghana there are a lot of surveys, which can be quite variable but generally follow monotonic trends. We can also see that sometimes the urban and rural trends are similar (e.g. wood), but they can differ (e.g. charcoal). There is also one potential outlier value for coal use, which may impact any modelling efforts (as will be discussed in Section 3.4.1). For Mongolia,

there are much fewer surveys. While they seem to follow smooth trends, it is difficult to know whether they reflect well the underlying trends, or if survey variability is hidden by the low number of surveys.

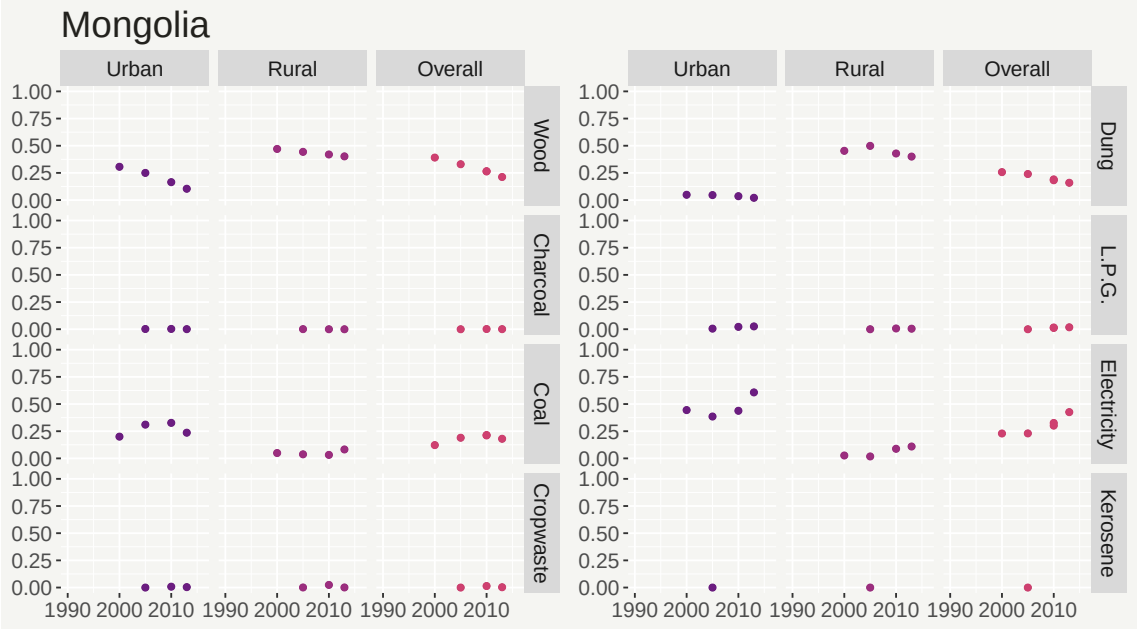

Figure 3.2: Time series of fuel use from surveys of Mongolia, in a 2018 snapshot of the WHO Household Energy Database.

While survey coverage is increasing, there are still many countries with an insufficient number of surveys to directly produce reliable estimates. To address this, statistical models can be used to pool information from other sources, such as covariates, in order to allow for more reliable inference in countries with insufficient survey data. For example, Rehfuess et al. (2006) use regression methods to quantify the association between solid fuel usage and a number of socio-economic factors, in order to predict usage in countries where no data was available.

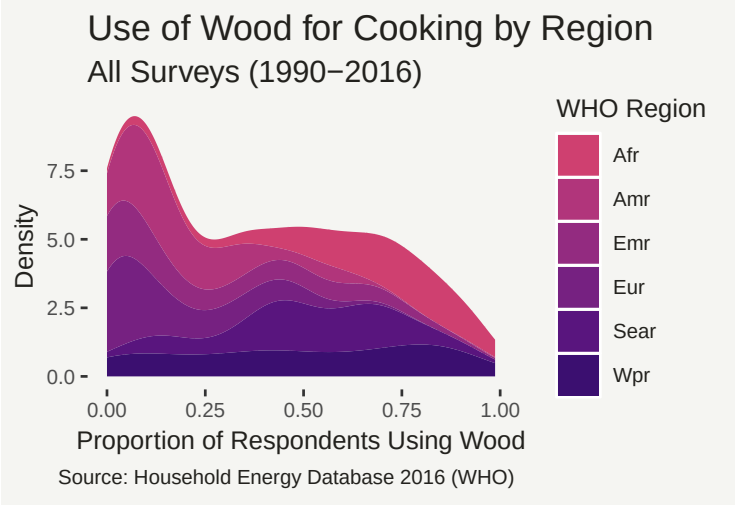

Figure 3.3: Smooth density estimates of the regional variability in the usage of wood as the primary cooking fuel by WHO region, and globally. Regions are AMR = Americas, EMR = Eastern Mediterranean, EUR = Europe, SEAR = South East Asia, WPR = Western Pacific.

An alternative source of information which can be exploited by statistical models is within-region similarity in cooking fuel use. Figure 3.3 illustrates regional differ-

ences by showing smooth density estimates of the distribution of the proportion of households using wood as the primary cooking fuel by different WHO regions. Regional pooling was adopted by Bonjour et al. (2013), using multi-level models with regional hierarchies to model trends in solid fuel usage, for the purpose of estimating the number of people exposed to household air pollution globally.

Previous approaches using data from the WHO Household Energy Database have focussed on understanding the proportion of population that use solid fuels as a group, rather than individual fuel usage. While a useful indicator for exposure to household air pollution, the limitation of only estimating the proportion of solid fuel usage in each country is that it inhibits interventions based on specific fuels, such as the deployment of cleaner wood burning stoves, while also failing to take into account the varying levels of harm caused by different fuels. Jointly modelling multivariate proportion data (e.g. from individual fuel types) is a challenging statistical problem in its own right, however there are also three notable flaws of the data which pose additional challenges:

(i) Many surveys only report values for a subset of the fuels of interest, meaning that the model must allow for surveys of varying levels of completeness. This ensures that inference regarding the use of the other fuels is possible on the basis of partial information and pooling across the data.

(ii) The total number of respondents is only available for approximately 50% of surveys, which prohibits the direct modelling of the number of respondents that use each fuel (e.g. with a Multinomial distribution). Instead, proportions (for which data is available) may be modelled using a multivariate distribution on the simplex, such as the Dirichlet. However, this would assume that individual values are strictly between 0 and 1, whereas there are numerous observations of 0% and 100% in the data considered here, making this impractical.

(iii) Although most surveys provide distinct observations for fuel use in urban and rural areas, a large number only provide information for the entire population (both urban and rural). The model must be applicable for both types and, in the case of surveys with only values for overall, allow estimation of the urban and rural values based on information on the proportion of people living in urban and rural areas (in that country). This is made more problematic as, for some countries, surveys include too many or too few urban or rural respondents, which introduces bias in the overall value. To allow for this, the model should also be able to estimate any systematic bias in the proportion of urban and rural people included in surveys, compared to external data on the proportions of the wider population living in urban and rural areas.

3.3 Model Design

For clarity of exposition, the following explanation relates to y_i, the number of respondents in a survey using fuel i as their primary fuel for cooking, ignoring for now any indices related to the country and the year. Here, $i = 1, \ldots, 9$ corresponds to wood, charcoal, coal, crop waste, dung, electricity, l.p.g., kerosene and finally an aggregation of other fuels (e.g. natural gas), which mostly constitute a very small percentage of the total. If we knew the total number of survey respondents n for all data, a first approach to modelling could be to assume that data on $\boldsymbol{y} = \{y_i\}$ arise from a Multinomial$(\boldsymbol{p}, n)$ distribution. Then, p_i would represent the proportion of people in the population using fuel i. This assumes that the survey sample is representative of the overall population. In reality, survey samples are imperfect and the Multinomial model may not be sufficiently flexible to capture the extra variability caused by flaws in the survey design. For instance, the survey may not cover the whole geographical area of interest.

A flexible extension of this approach is to model $\boldsymbol{y}$ using a Generalized-Dirichlet-Multinomial$(\boldsymbol{\alpha}, \boldsymbol{\beta}, n)$ (GDM) distribution, a mixture of the Generalized-Dirichlet with pdf:

$$p(p_1, p_2, ..., p_k \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p_k^{\beta_{k-1}-1} \prod_{i=1}^{k-1} \left[\frac{p_i^{\alpha_i - 1}}{B(\alpha_i, \beta_i)} \left(\sum_{j=i}^{k} p_j \right)^{\beta_{i-1} - (\alpha_i + \beta_i)} \right] \tag{3.1}$$

and the Multinomial distribution, so that

$$\boldsymbol{p} \sim \text{Generalized-Dirichlet}(\boldsymbol{\alpha}, \boldsymbol{\beta}); \qquad \boldsymbol{y} \mid \boldsymbol{p} \sim \text{Multinomial}(\boldsymbol{p}, n) \tag{3.2}$$

with pdf:

$$p(y_1, y_2, ..., y_k \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, n) = \frac{\Gamma(n+1)}{\Gamma(y_k + 1)} \prod_{i=1}^{k-1} \left[\frac{\Gamma(y_i + \alpha_i)\Gamma(\sum_{j=i+1}^{k} y_j + \beta_i)}{B(\alpha_i, \beta_i)\Gamma(y_i + 1)\Gamma(\alpha_i + \beta_i + \sum_{j=i}^{k} y_j)} \right] \tag{3.3}$$

This means that any additional variability caused by flawed sampling can be potentially captured by the Generalized-Dirichlet component. The Generalized-Dirichlet also has a very flexible covariance structure compared to the Dirichlet, which it reduces to in the special case that $\beta_i = \alpha_{i+1} + \beta_{i+1}$ for $i \in 1, ..., k-2$ and $\beta_{k-1} = \alpha_k$.

Recall that for most of the available data, only the proportion $\boldsymbol{x} = \{y_i/n\}$ of respondents using each fuel is available, with the total number of respondents n being unknown. This means that the GDM cannot be used to directly model the number of respondents primarily using each fuel, if one wishes to use all of the available data. However, here the principal interest lies in estimating the fuel usage proportions $\boldsymbol{x}$, so an alternative approach would be to model the proportions themselves, for example using a Generalized-Dirichlet distribution. In that case, though, the presence of many 0% and 100% fuel usage observations (which fall outside the

range space of the Generalised-Dirichlet) make this impractical. Instead, we opt for an approximate procedure for modelling $\boldsymbol{x}$, namely by transforming observations of x_i into conceptual counts v_i, out of a user-chosen total N. To ensure that the sum of the transformed counts does not exceed N, one can compute $v_i = \lfloor N x_i \rfloor$ (as opposed to rounding). The counts $\boldsymbol{v}$ can then be modelled as GDM$(\boldsymbol{\alpha}, \boldsymbol{\beta}, N)$ so that predictions are based on v_i/N.

Using this method, we can obtain approximately the same inference for the underlying usage in the wider population as if we had modelled $\boldsymbol{y}$ directly. In addition, the flexibility of the GDM means that we can still capture the distribution of $\boldsymbol{x}$ well. This is because any variability lost or gained from the Multinomial component, by respectively using a larger or a smaller N compared to the original n, can be accounted for by appropriate adjustment in the parameters of the underlying Generalized-Dirichlet component.

3.3.1 Simulation experiment

To illustrate the validity of our approximation for modelling the proportions using each fuel type $\boldsymbol{x} = \boldsymbol{y}/n$, as well as to assist in our choice of N, we present a simulation experiment using the 598 observed survey samples sizes n. The majority are in the range 1000-100000, with a mode of around 10000. At these large values, the contribution of the Multinomial variance to the total variance of $\boldsymbol{x}$ is quite small.

For each available n_i $(i = 1, \ldots, 598)$, we simulate a vector of survey responses $\boldsymbol{y}_i = \{y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4}\}$ from a GDM model. Here, each country has a different (time constant) marginal mean vector $\boldsymbol{\mu}_c$ and variance parameters $\boldsymbol{\phi}_c$ (preserving the original associations between the countries and observed n_i in the data, and ignoring countries with no observed n_i) . Note that some countries will only have one $\boldsymbol{y}_i$ and others will have several (each with its own unique n_i). We simulate all of the $\boldsymbol{\mu}_c$ from a Dirichlet$(\mathbf{1})$ distribution, and all of the $\boldsymbol{\phi}_c$ independently from a Gamma$(4, 0.1)$ distribution (inducing a moderately high degree of over-dispersion, compared to the Multinomial):

$$\boldsymbol{y}_i \sim \text{GDM}(\boldsymbol{\mu}_c, \boldsymbol{\phi}_c, n_i) \tag{3.4}$$

$$\boldsymbol{\mu}_c \sim \text{Dirichlet}(\mathbf{1}) \tag{3.5}$$

$$\boldsymbol{\phi}_c \sim \text{Gamma}(4, 0.1) \tag{3.6}$$

In the baseline scenario, to which we will compare our approximate method, we have observations for all of the n_i and all of the $\boldsymbol{y}_i$. This allows us to implement the above model directly, which we do in a Bayesian setting using a Dirichlet$(\mathbf{1})$ prior for each $\boldsymbol{\mu}_c$ and a non-informative Exponential(0.001) prior for each ϕ_c.

In the second scenario, we don't know any of the n_i or the $\boldsymbol{y}_i$, but we do have observations for $\boldsymbol{x}_i = \boldsymbol{y}_i/n_i$. In this scenario, we can apply our approximate method

(from Section 3.3), where we fit the GDM to constructed counts $\boldsymbol{v}_i = \lfloor N\boldsymbol{x}_i \rfloor$. We proceed to apply this method whilst varying N over a range of values (10, 20, 30, 50, 100, 300, 1000, 3000, 10000, 30000, 100000, 300000, 1000000), so that we can investigate the impact of this choice on parameter inference.

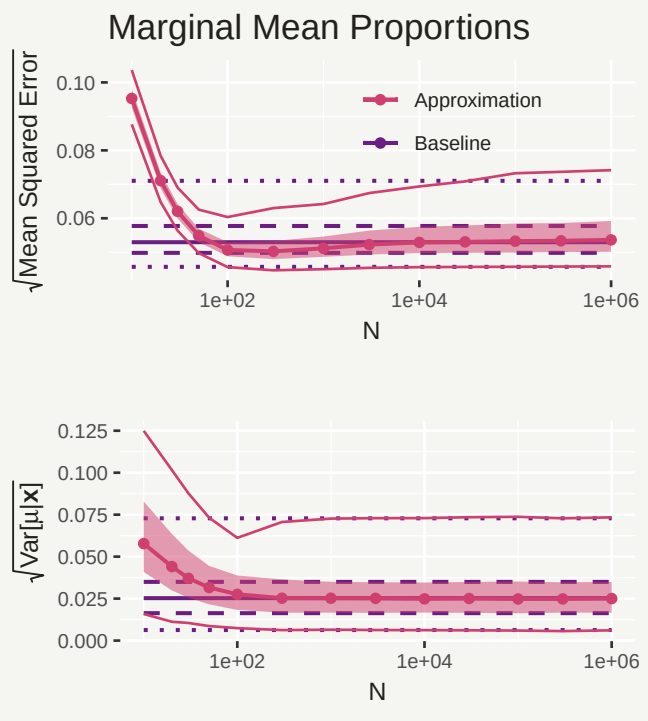

Figure 3.4: The top panel shows the median, interquartile range (dark) and 95% interval (light) of the mean squared differences between the posterior samples of the marginal mean proportions $\mu_{1,c}, \ldots, \mu_{4,c}$ and their corresponding true values, from the approximate model with varying N. Similarly, the bottom plot shows the median, interquartile range and 95% interval of the posterior standard deviations of $\mu_{1,c}, \ldots, \mu_{4,c}$. The dashed lines represent these results from the baseline model.

Recall that in our application we are primarily interested in correct inference for the marginal mean proportions $\boldsymbol{\mu}_c$ (the underlying fuel use in each country), and we claimed that a sufficiently large choice of N yields a parameter inference approximately the same as if we had modelled the $\boldsymbol{y}_i$ directly, along with the sample sizes n_i. To assess this, we begin by examining the models' accuracy when predicting the true marginal mean proportions $\boldsymbol{\mu}_c$. For each posterior sample, we can compute the mean squared error between the predicted values of $\boldsymbol{\mu}_c$ and the true values. The top panel of Figure 3.4 shows the median of this statistic, for varying N, as well as the inter-quartile range (dark), and 95% prediction interval (light). Compared to the same statistics for the baseline model, shown as horizontal lines, we can see that the distribution of mean squared errors for the approximate method does indeed appear to converge to the baseline model as N increases, from about $N = 10000$ onwards.

We can also examine how the approximate method quantifies uncertainty in $\boldsymbol{\mu}_c$. For each individual $\mu_{1,c}, \ldots, \mu_{4,c}$, we compute the standard deviation of the posterior samples. The median of these posterior samples are then shown for each N in the bottom panel of Figure 3.4, once again alongside the inter-quartile range and 95% interval. The distribution of posterior standard deviations for the approximate method also converges to the baseline model, but does so for a much lower N (between 100 and 1000) than the mean squared error.

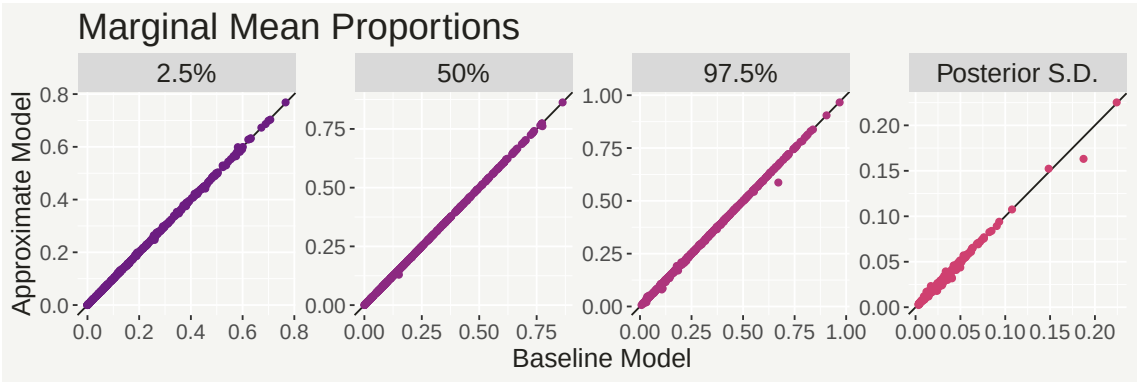

Figure 3.5: Scatter plots comparing the posterior 2.5% (left), 50% (second from left), and 97.5% (second from right) posterior quantiles, and posterior standard deviations (right) for the marginal mean proportions $\mu_{1,c}, \ldots, \mu_{4,c}$, from the approximate model with $N = 10000$, to the baseline model.

Finally, if we choose a single value of N, we can compare more closely the approximate method to the baseline model when estimating $\boldsymbol{\mu}_c$. Figure 3.5 compares the 2.5% (left), 50% (second from left), and 97.5% (second from right) posterior quantiles for the $\mu_{1,c}, \ldots, \mu_{4,c}$ from the approximate model with $N = 10000$, to the sample quantiles from the baseline model. The quantiles are virtually identical, suggesting that for this simulated data the inference would be achieved either by modelling the true counts $\boldsymbol{y}_i$ directly or by modelling the constructed counts $\boldsymbol{v}_i = \lfloor 10000 * \boldsymbol{x}_i \rfloor$.

The results in this thesis were generated by opting for $N = 1000$, but in subsequent versions of the model we have instead used a more conservative value $N = 100000$. Empirically, however, we have not noticed any difference in the results when changing between these two values.

3.3.2 Conditional models

Recall that one of the main issues with the given data is that quite often, a value x_i (and thus v_i) for at least one individual fuel is missing (for a given country-year combination). To model this data in a way that inference is made possible on the missing fuel proportions, it makes sense to implement the GDM using the implicit conditional densities rather than the joint one. Specifically, for counts $\boldsymbol{v}$ and total N, the conditional distribution of each (fuel) v_i given the others (and parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) is:

$$v_1 \sim \text{Beta-Binomial}\left(\alpha_1, \beta_1, n_1 = N - v_1\right) \tag{3.7}$$

$$v_i \mid v_1, \ldots, v_{i-1} \sim \text{Beta-Binomial}\left(\alpha_i, \beta_i, n_i = N - \sum_{j=1}^{i-1} v_j\right) \tag{3.8}$$

$$p(v_i \mid v_1, \ldots, v_{i-1}) = \binom{n_i}{v_i} \frac{B(v_i + \alpha_i, n_i - v_i + \beta_i)}{B(\alpha_i, \beta_i)} \tag{3.9}$$

for $i = 2, ..., k$. Fitting this model in a Bayesian setting implies that any missing values v_i can be sampled using Markov Chain Monte Carlo (MCMC).

For ease of interpretation, it makes sense to re-parameterize the conditional distributions in terms of their expectations ν_i and variance parameters ϕ_i:

$$\alpha_i = \nu_i \phi_i; \qquad \beta_i = (1 - \nu_i)\phi_i \tag{3.10}$$

The relative mean ν_i is interpreted as the mean proportion of households using fuel i out of those not using any of the fuels higher up the hierarchy $(1, \ldots, i-1)$. For example, ν_1 is the underlying proportion who use wood from the whole population, ν_2 is the proportion who use charcoal from the population who do not use wood and ν_3 is the proportion who use coal from the population who use neither wood nor charcoal. Through parameter ϕ_i, the model is able to compensate for any gain or loss of variance in the conditional Multinomial component caused by the introduction of the "artificial" total N.

It is noted that for the SDG indicators, the primary quantity of interest is the marginal mean vector of proportions $\boldsymbol{\mu} = \{\mu_i\}$ of households relying on each fuel i. This can be recovered from the relative means ν_i:

$$\begin{aligned} \mu_1 &= \nu_1 && (3.11)\\ \mu_2 &= \nu_2(1 - \nu_1) && (3.12)\\ &\vdots \\ \mu_k &= \nu_k \prod_{i=1}^{k-1}(1 - \nu_i) && (3.13) \end{aligned}$$

Now introducing indices for a survey conducted in WHO region r, country c and year t, the characterisation of the relative mean $\nu_{i,r,c,t}$ is defined by:

$$\log\left(\frac{\nu_{i,r,c,t}}{1 - \nu_{i,r,c,t}}\right) = \gamma_{i,r,1} + \delta_{i,c,1} + (\gamma_{i,r,2} + \delta_{i,c,2})t \tag{3.14}$$

where the logistic transformation ensures that $\nu_{i,r,c,t} \in (0, 1)$. Fixed effects $\boldsymbol{\gamma}_{i,r} = (\gamma_{i,r,1}, \gamma_{i,r,2})$ capture regional (linear) changes over time in mean fuel use. Constrained (random) effects $\delta_{i,c,1}$ and $\delta_{i,c,2}$ allow countries to deviate from the regional trend, if there is sufficient evidence in the data that they should. We return to the model for the variance parameters ϕ_i later on.

3.3.3 Rural and urban variability

A further source of variability in fuel usage arises from differences between rural and urban areas. The model captures this by allowing the regional trends as well as the country differences in (3.14) to be different for rural/urban areas. It is likely that these differences within a country are correlated, so to capture this we model each

pair $\boldsymbol{\delta}_{i,c,j} = (\delta^{urban}_{i,c,j}, \delta^{rural}_{i,c,j})$ with a Multivariate-Normal distribution:

$$\begin{pmatrix} \delta^{urban}_{i,c,j} \\ \delta^{rural}_{i,c,j} \end{pmatrix} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}^{\delta}_{i,r,j}) \text{ for } j = 1, 2 \tag{3.15}$$

where the covariance matrix $\boldsymbol{\Sigma}^{\delta}_{i,r,j}$ is allowed to differ between regions. Unfortunately, while most surveys in the data report both urban and rural values, this is not always the case. Some only report an overall value for the whole sample. So that we can still use this information, we incorporate a layer in the model to constrain the marginal mean proportions as follows:

$$\boldsymbol{\mu}^{overall}_{r,c,t} = \pi_{c,t}\boldsymbol{\mu}^{urban}_{r,c,t} + (1 - \pi_{c,t})\boldsymbol{\mu}^{rural}_{r,c,t} \tag{3.16}$$

$$\log\left(\frac{\pi_{c,t}}{1 - \pi_{c,t}}\right) = \log\left(\frac{P_{c,t}}{1 - P_{c,t}}\right) + f_c(t) \tag{3.17}$$

The overall mean proportion of fuel usage $\boldsymbol{\mu}^{overall}_{r,c,t}$ is a vector of individual fuel usage, as defined in (3.11)–(3.13), for each region r, country c and year t. This is then defined as a weighted sum in (3.16), of the rural and urban mean proportions. The weights $\pi_{c,t} \in (0, 1)$ represent the mean proportion of survey respondents living in an urban area, in country c and year t. Furthermore, $P_{c,t}$ are estimates of the proportion of people living in an urban area for each country and year (United Nations, 2018) and they are used as offsets in a model for $\pi_{c,t}$. For each country, systematic deviations from these estimates are modelled using a smooth function $f_c(t)$, to allow for potential under- or over-sampling of urban populations in the survey data. In terms of the modular framework for accounting for flawed observation mechanisms discussed in Chapter 1, we can think of the model for $\boldsymbol{\mu}^{urban}_{r,c,t}$ and $\boldsymbol{\mu}^{rural}_{r,c,t}$ as the latent model for the quantity we are interested in (the underlying usage of each fuel), and the inclusion of $f_c(t)$ in (3.17) as a flawed observation module to account for sampling bias. Since $f_c(t)$ is a correcting factor, it should ideally be flexible enough to capture any systematic sampling biases with respect to the UN estimates $P_{c,t}$. However, from a modelling perspective, this introduces extra degrees of freedom for the model to capture the overall survey observations well. Therefore, to avoid over-fitting, we model $f_c(t)$ using penalised low-rank thin-plate splines (Crainiceanu et al., 2005), using a different smoothing penalty parameter σ_c for each country. This allows $f_c(t)$ to capture non-linear deviations from $P_{c,t}$ over time, but only when there is ample evidence of non-linearity in the data for a given country.

We complete the model specification by defining the following structure in the variance parameters $\phi_{i,r,c}$:

$$\log(\phi_{i,r,c}) = \psi_{i,r} + \epsilon_{i,c} \tag{3.18}$$

where $\psi_{i,r}$ are fixed regional effects and $\epsilon_{i,c}$ are country-level deviations from these. This allows for the fact that average survey size and representativeness can vary

between countries, affecting the variance of the observed fuel proportions. As with the means, we believe the variance parameters for urban, rural and overall survey values to be correlated within a country, and so to capture this we use a Multivariate-Normal model for the triples $\boldsymbol{\epsilon}_{i,c} = (\epsilon_{i,c}^{urban}, \epsilon_{i,c}^{rural}, \epsilon_{i,c}^{overall})$:

$$\begin{pmatrix} \epsilon_{i,c}^{urban} \\ \epsilon_{i,c}^{rural} \\ \epsilon_{i,c}^{overall} \end{pmatrix} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}_{i,r}^{\epsilon}) \tag{3.19}$$

Such that, as with the model for the $\boldsymbol{\delta}_{i,c,j}$, the covariance structure is allowed to differ between regions.

3.3.4 Prior distributions

For the regional (fixed) effects $\boldsymbol{\gamma}_{i,r}$ and $\psi_{i,r}$, we chose to specify weakly informative Normal$(0, 10^2)$ and Normal$(0, 2^2)$ prior distributions, respectively. These reflect our limited prior knowledge of regional trends. For more efficient MCMC sampling, the country effects $\boldsymbol{\delta}_{i,c,j}$ and $\boldsymbol{\epsilon}_{i,c}$ were centred on the regional effects, so that their Multivariate-Normal means were (instead of $\mathbf{0}$ as in (3.15) and (3.19)) $\boldsymbol{\gamma}_{i,r,j}$ and $\boldsymbol{\psi}_{i,r}$, respectively, with:

$$\log\left(\frac{\boldsymbol{\nu}_{i,r,c,t}}{1-\boldsymbol{\nu}_{i,r,c,t}}\right) = \boldsymbol{\delta}_{i,c,1} + \boldsymbol{\delta}_{i,c,2}t \tag{3.20}$$

$$\log(\boldsymbol{\phi}_{i,r,c}) = \boldsymbol{\epsilon}_{i,c} \tag{3.21}$$

For the Multivariate-Normal covariance matrices, we chose conjugate Inverse-Wishart prior distributions with informative marginal distributions for the variances and correlations, with more prior density over positive correlations than negative correlations. For the smoothing penalty parameters σ_c we assigned Half-Normal$(0, 1^2)$ prior distributions. This reflects our belief that smaller values for σ_c (corresponding to a stronger penalty) are more likely than larger ones.

3.3.5 Survey Selection

The model was applied to a selection of the February 2018 version of the WHO Household Energy Database. Surveys were excluded from the analyses if:

- They only reported the usage of 'solid fuels' as a group, rather than the usage of at least one individual fuel.
- They included a high proportion (>15%) of respondents who either reported that they cook with an unlisted fuel, do not cook at all or who failed to respond. These surveys were deemed to be excessively 'incomplete' and were not included for modelling. This threshold was chosen subject to sensitivity analysis.

Surveys removed for exceeding the incompleteness threshold are shown as black points in the Supplementary Material.

3.3.6 Implementation

All code was written and executed using R (R Core Team (2018)) and the model was implemented using NIMBLE (de Valpine et al., 2017), a facility for highly flexible implementation of Markov Chain Monte Carlo (MCMC) models. For this application, we needed to add the Beta-Binomial distribution to NIMBLE, which was straightforward using only a few lines of R code. Four MCMC chains were run for 80,000 iterations from different randomly generated initial values and with different random number generator seeds. The first 40,000 samples were then discarded as burn-in and, to limit system memory usage, the remaining samples were thinned by 10.

Assessing the convergence of the MCMC chains is made challenging by the extremely high number of parameters (tens of thousands) in the model. Recall that the intercept and slope effects $\boldsymbol{\delta}_{i,c,j}$ and variance effects $\boldsymbol{\epsilon}_{i,c}$ were centred around the regional fixed effects. This means that the fuel usage means $\boldsymbol{\mu}_{i,r,c,t}$ are completely defined by $\boldsymbol{\delta}_{i,c,j}$ (and in the case of overall fuel usage, the urban proportions $\pi_{c,t}$) and that the variance parameters $\boldsymbol{\phi}_{i,r,c}$ are completely defined by $\boldsymbol{\epsilon}_{i,c}$. Therefore we can assess the convergence of the model by assessing the convergence of $\boldsymbol{\delta}_{i,c,j}$, $\pi_{c,t}$ and $\boldsymbol{\epsilon}_{i,c}$.

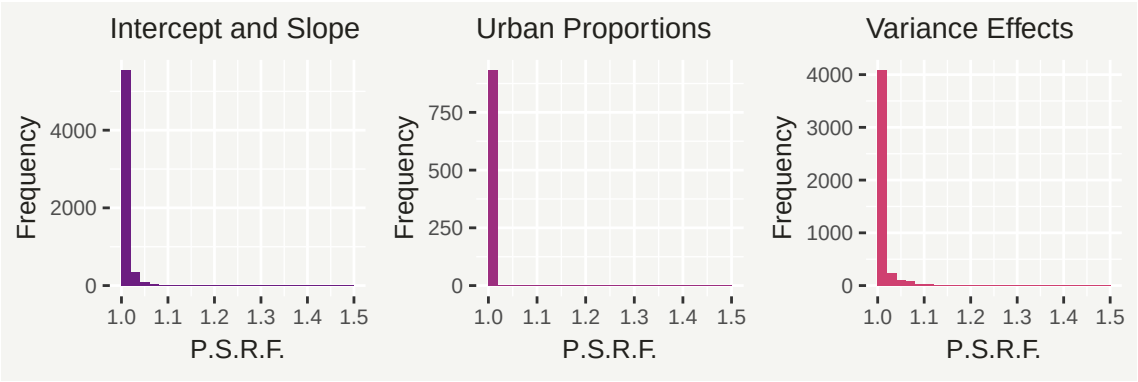

Figure 3.6: Histograms of the Potential Scale Reduction Factor (PSRF) computed for the intercept and slope effects $\boldsymbol{\delta}_{i,c,j}$ (left), the urban proportions $\pi_{c,t}$ (centre) and the variance effects $\boldsymbol{\epsilon}_{i,c}$ (right).

To do this, we computed the PSRF (detailed in Section 2.3.4) for the (6208) intercept and slope effects $\boldsymbol{\delta}_{i,c,j}$, the (993) urban proportions $\pi_{c,t}$ and the (4656) variance effects $\boldsymbol{\epsilon}_{i,c}$ and Figure 3.6 presents them respectively in frequency histograms. For all three sets of parameters, the overwhelming majority of the values lie in the closest bin to 1, suggesting that the model has converged. All the associated code required to implement the model is available in the Supplementary Material.

3.4 Model Checking

The task of assessing the validity of the statistical model is divided into two parts. The first comprises basic procedures to check that the model has no systematic issues with reproducing the observed data, while the second assesses the ability of the model to predict future fuel usage values.

3.4.1 Posterior predictive checking

Given the Bayesian implementation of the model, assessing the fit to both in-sample and out-of-sample data is based on posterior predictive model checking (Gelman et al., 2014). For in-sample data, this involves simulating from the posterior distribution of parameters and random effects (samples of which are already available from MCMC) and then simulating v_i from the conditional Multinomial distribution to obtain samples of the posterior predictive distribution for replicates $\tilde{\boldsymbol{x}}|\boldsymbol{x}$ of the observed fuel proportions $\boldsymbol{x}$. The statistical properties of these replicates can then be compared to properties of the corresponding observations.

In the first instance, scatter plots comparing the posterior means of the replicates with the observed values can give an indication of any systematic issues. Figure 3.7 shows scatter plots comparing the mean predicted replicates for wood, charcoal, crop waste and coal to their corresponding observed values and Figure 3.8 shows the same plots for dung, electricity, l.p.g. and kerosene. In general the points are scattered about the diagonal line fairly evenly, indicating a good model fit for the different fuels. However, there are some patterns among the different fuels worth discussing. First, the variation around the diagonal differs considerably between the fuels. For example, the differences between the predictions and observed values are more variable for l.p.g. than for wood. This suggests that the model may be more precise when predicting wood usage than l.p.g., though in both cases the coverage of the 95% intervals is very high.

Additionally, there is some notable systematic deviation from the diagonal in the plot of overall electricity values, where a string of observed values exceeds the mean predicted replicates. Upon closer inspection, this was found to be the survey values for electricity in South Africa, where the model is distorted by a single outlier (as discussed in Section 3.4.1).

Also shown are coverage values, the proportion of observed values which lie within the 95% posterior predictive intervals of the corresponding replicates. A coverage substantially lower than 95% would mean a high proportion of observed values are extreme values with respect to the posterior predictive model, implying a poor fit. In this case, the coverage values for the 95% credible intervals were higher than 95% for all fuels and areas. Taken together, these two checks indicate that the model captures the observed data well.

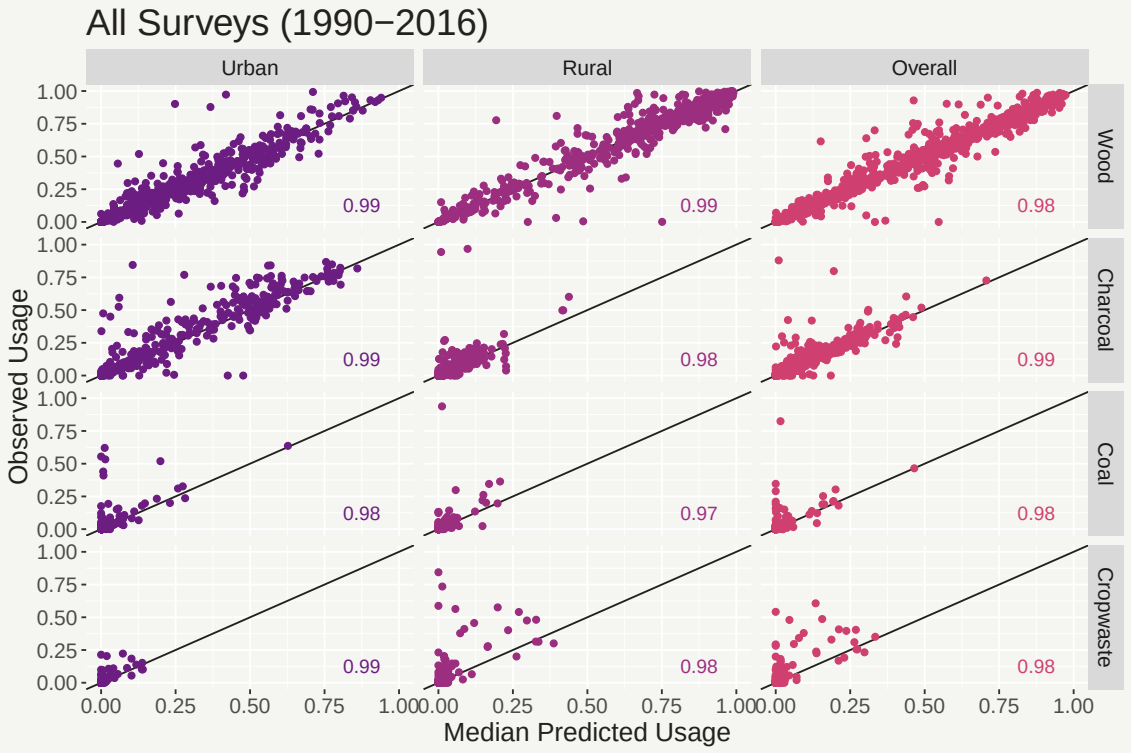

Figure 3.7: Scatter plots comparing the posterior means of wood, charcoal, cropwaste and coal usage replicates to their corresponding observed values.

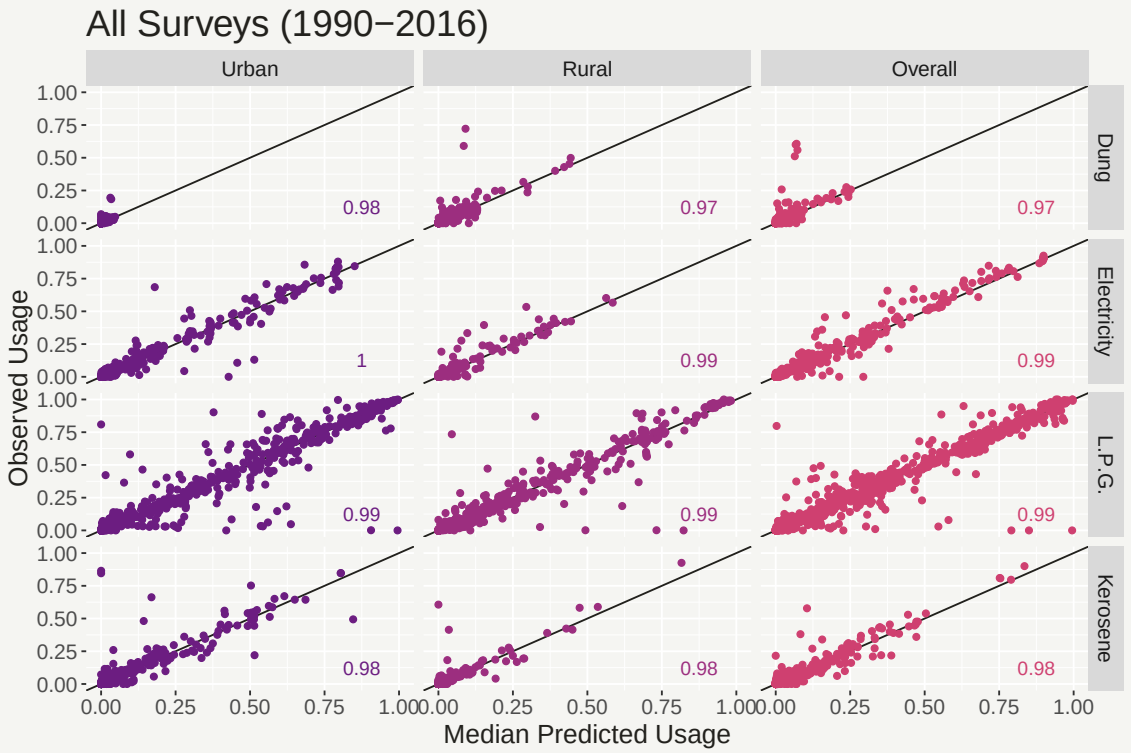

Figure 3.8: Scatter plots comparing the posterior means of dung, electricity, l.p.g. and kerosene usage replicates to their corresponding observed values.

Another way of checking the model is to compare predicted trends to survey observations on an individual country basis. Figure 3.9 shows the mean predicted trend for the proportion using each fuel in each segment (urban, rural and overall)

of Ethiopia, with associated 95% posterior predictive intervals. Here it can be seen that the mean trend lines follow the observed trends well, with prediction intervals that envelop a reasonable number of surveys. Moreover, by examining the tightness of the prediction intervals with respect to the variance of the observations, we can verify that the high coverage values obtained for the replicate prediction intervals are not simply caused by excessively high model uncertainty.

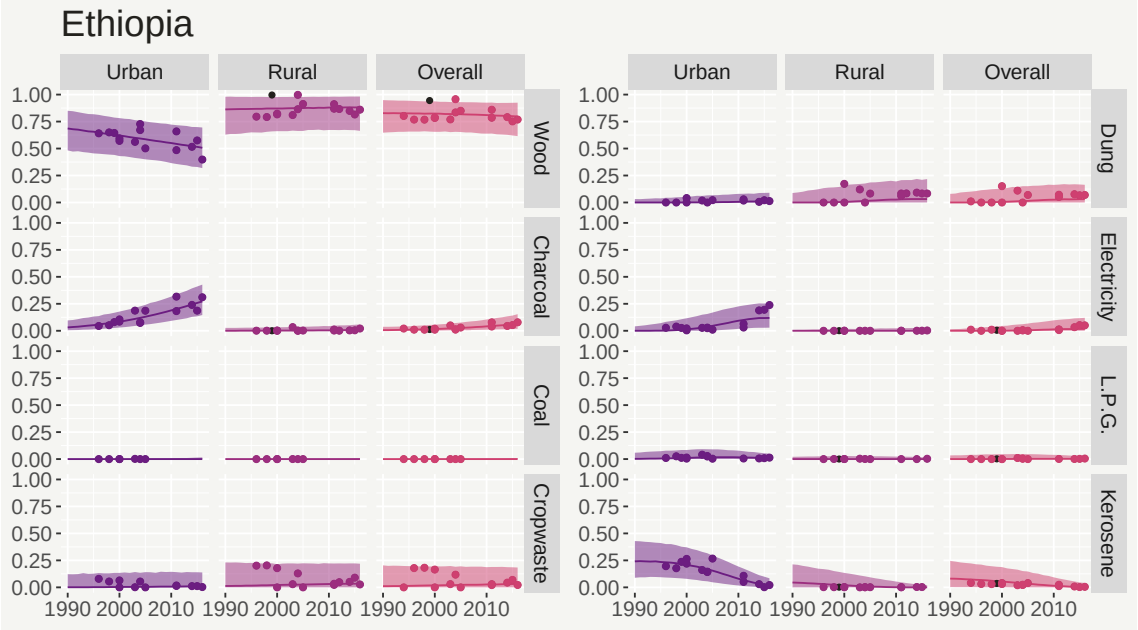

Figure 3.9: Mean predicted fuel usage trends with associated 95% posterior predictive intervals for Ethiopia. Coloured points represent survey observations and black points represent removed surveys. For each fuel, the left, central and right plots show urban, rural and overall usage, respectively.

The same plots can serve as a useful tool for identifying surveys which don't align with the general pattern in a given country. Figure 3.10 shows the predicted trends for South Africa and it can be seen that all surveys report substantial use of electricity for cooking, except for one survey which reports zero usage. The model has sought to capture the conflicting information with a mean trend line between the two, accompanied by wide prediction intervals. While not completely implausible, the plot indicates this may warrant further investigation.

Note that to check the model reproduces the observed data well, the overall predictions in Figures 3.9 and 3.10 incorporate the model's prediction of any systematic deviation from the UN estimates of urban and rural proportions, in the sampling of urban and rural respondents. Instead, predictions of overall fuel usage can be based solely on the UN estimates of urban and rural proportions (rather than based on the proportions in the surveys), which may constitute a more robust summary of fuel usage in a given country. This is achieved by removing $f_c(t)$ from (3.17) during simulation. Fuel usage plots for both the prediction of new surveys (which include sampling variation and potential sampling biases, as in Figures 3.9 and 3.10) and

the prediction of the underlying fuel usage in the population ($\boldsymbol{\mu}_{i,r,c,t}$) are available for all countries in the Supplementary Material.

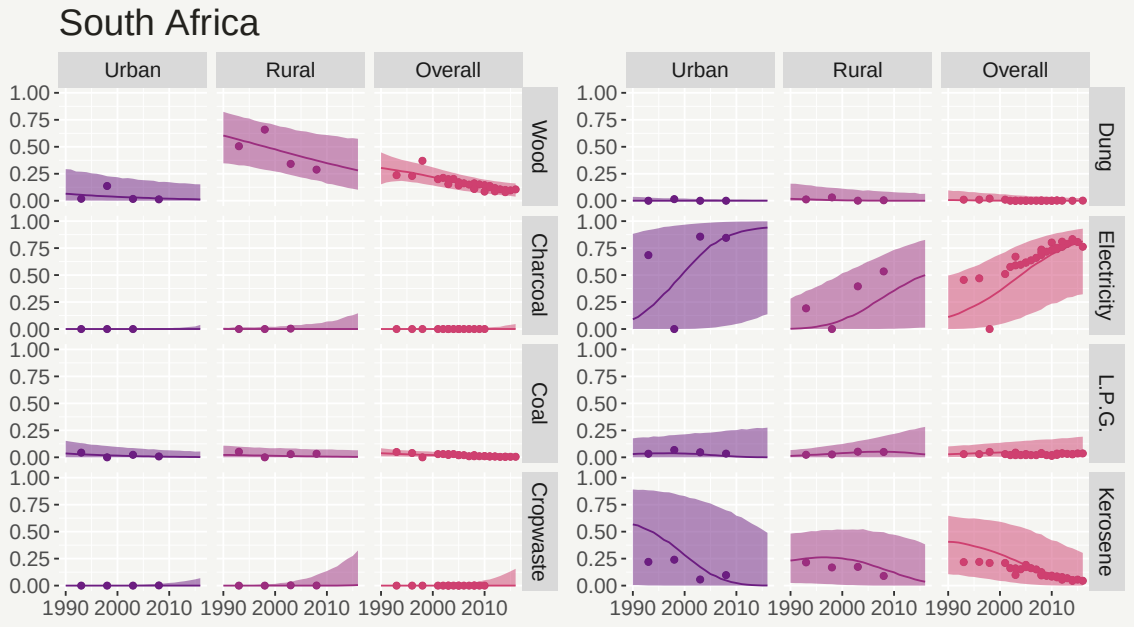

Figure 3.10: Mean predicted fuel usage trends with associated 95% posterior predictive intervals for South Africa.

The model's ability to capture systematic biases in the proportions of urban and rural respondents in the survey samples, relative to UN estimates, can also be inspected; Figure 3.11 shows the proportion of respondents recorded as urban in the fuel surveys for India (left) and Malawi (right) compared to UN estimates and predicted values from the model. The plot for India shows evidence of systematic over-sampling of urban respondents between 1997 and 2007, compared to the UN estimates. The plot for Malawi, meanwhile, shows limited evidence of any systematic deviation. It is likely that ignoring potential systematic biases would result in a less reliable inference for the relationship between urban, rural and overall surveys. In both of these cases, the spline incorporated in (3.17) appears to capture any differences well and therefore the associated bias can be mitigated when predicting and forecasting fuel usage.

3.4.2 Forecasting experiment

Samples from posterior predictive distributions for out-of-sample data are obtained in the same way as in-sample data, albeit using future time points as covariate values. The model's ability to predict (forecast) can be assessed using out-of-sample predictive testing. This is particularly important for this model to evaluate its suitability for forecasting future fuel use. To emulate a hypothetical forecasting scenario, the model was fitted only to surveys up to and including year 2012, therefore excluding 4 years (approximately one third of the data). We then used the model to predict 4

years into the future and produce predictive distributions for the out-of-sample data. Note that forecasting future overall survey observations involves forecasting how any systematic trends in the sampling of urban and rural respondents will progress in the future. While this may be possible it is not our primary interest, so we focus on checking the out-of-sample prediction of urban and rural surveys.

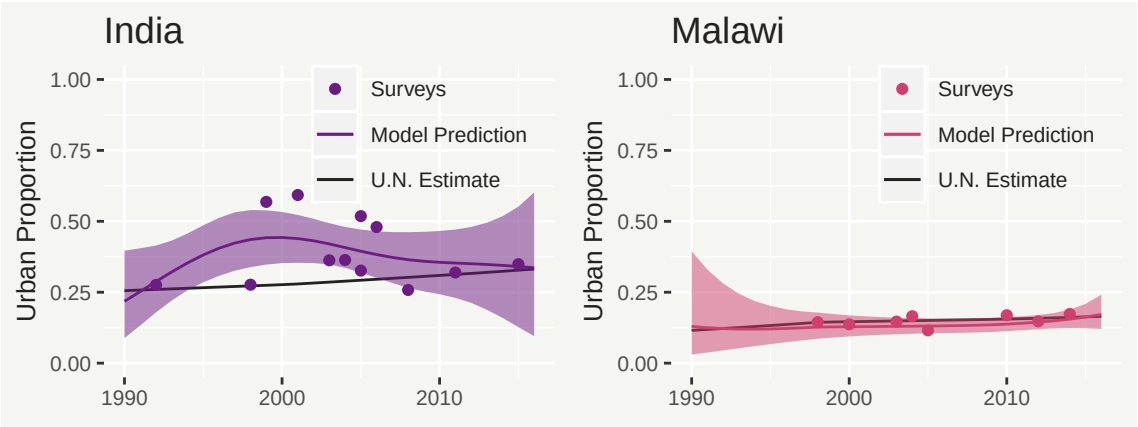

Figure 3.11: Plot of the urban proportions of fuel survey respondents in India (left) and Malawi (right), shown as points, compared to the U.N. estimates of the proportion of the respective populations living in an urban environment. Also plotted are the model's predictions of the change in each country's mean urban proportion of the surveys, with 95% credible intervals.

Figure 3.12 shows scatter plots comparing the out-of-sample survey values to the mean predicted values from the model. While there are some values which are not captured well (some potentially due to erroneous data), generally the model does not seem to systematically over or under-predict. The only exception to this is crop waste (urban), where there does appear to be some systematic deviation from the diagonal. Notably, the coverage values tend to be quite high, indicating that the model produces reliable uncertainty estimates when predicting into the future.

To guard against high coverage values through unreasonably uncertain prediction intervals, we can assess the model's performance when forecasting by examining predictive plots for individual countries. Figure 3.13 shows predictive fuel usage plots for Nepal, from the model where surveys from 2013 onwards are excluded. Though for some fuels the excluded surveys deviate from the mean predicted trend, they are generally well within the 95% predictive intervals, which are not so wide that they are impractical. Looking at urban areas, we can see that the surveys up to 2012 suggest the use of LPG might continue to increase from 2013 onwards, although the model appears to correctly predict the plateauing that is present in the results from the excluded surveys.

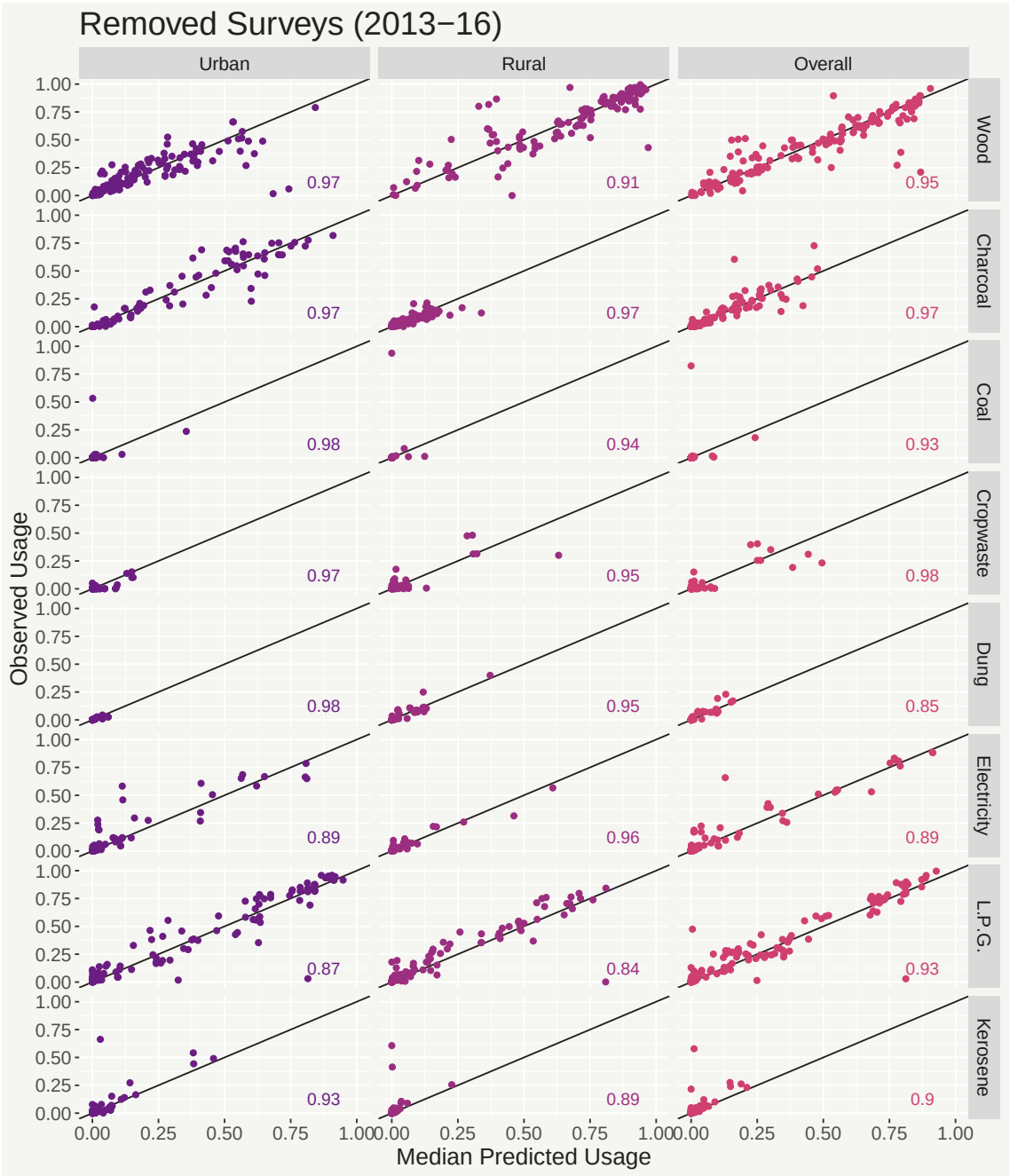

Figure 3.12: Scatter plots of mean predicted fuel usage values from 2013 onwards, versus their observed values, from the model which was only supplied data from 2012 or earlier.

3.5 Discussion

In this chapter, we have developed a multivariate hierarchical model, based on Generalized-Dirichlet-Multinomial distributions, to model trends in the use of polluting and clean fuels for cooking across the world. The work was motivated by the need to expand the evidence base related to household use of individual fuels that is crucial when developing policy and planning interventions. The principal aim was to estimate changes in the use of individual fuels within the period 1990-2016. This was achieved for each country, with distinction between urban and rural areas,

together with predictions of future fuel usage. The proposed approach addresses the inherent difficulty in jointly modelling multivariate proportion data, and several other challenges associated with modelling the data from the WHO Household Energy Database. These challenges included missing values for the use of some fuels within surveys, the total number of respondents only being available for approximately half of all surveys, some surveys not distinguishing between urban and rural areas, and biases in the sampling of urban and rural respondents.

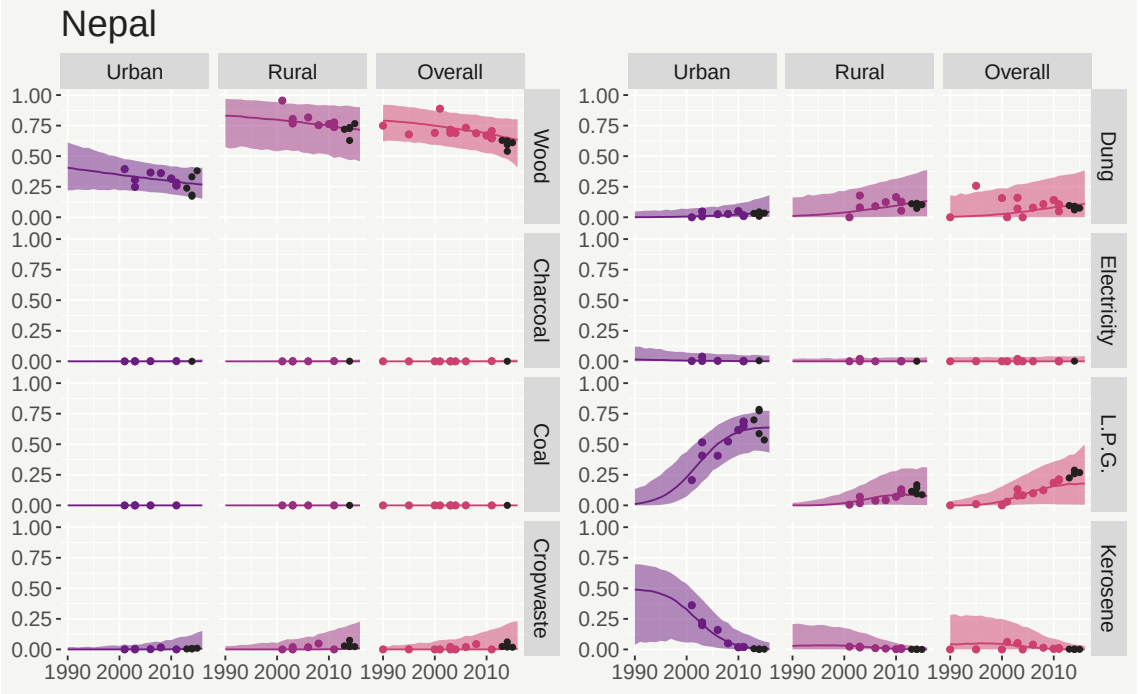

Figure 3.13: Mean predicted fuel usage trends with associated 95% posterior predictive intervals for Nepal, from the model where surveys from 2013 onwards were excluded. The black points from 2013 onwards represent values from excluded surveys.

The resulting global household energy model (GHEM) is implemented within a Bayesian hierarhical framework. Trends in the proportions of populations using each fuel are estimated for each country, based primarily on information from surveys within that country. Where data are not available within a country, or are insufficient to produce accurate estimates, the model structure 'borrows' information from regional trends and, in such cases, the associated uncerainty is increased. The model also takes into account, and estimates, any systematic biases in the sampling of urban and rural respondents. The primary output of the model is the underlying fuel usage in the sampled population, represented by the $\boldsymbol{\mu}_{r,c,t}$ in Equation (3.16). This constitutes a more robust and stable measure on which to base policy decisions than using individual surveys.

Predicting future patterns of fuel usage from the model using the estimated trends provides a baseline representation of what might be expected in the absence of intervention and provides a comparison against which future surveys conducted

post-interventions could be compared. The advantage of modelling the relative fuel means ($\boldsymbol{\nu}_{r,c,t}$, in Equation (3.14)) as linear in time (on the logistic scale) is that it is possible to extrapolate observed trends arbitrarily far into the future. However, this should be done in the context of the forecasting experiment in Section 3.4.2, which suggests forecasts might be restricted to a few years into the future, beyond which it is possible that the logistic-linear approximation may not be reasonable.

The model has been used by the WHO to produce estimates of the number of people in each country who use polluting fuels for cooking, to provide proxy for exposure to household air pollution and to assess the take-up of clean fuel technologies (World Health Organization, 2018c). During its development, the model has played, and will continue to play, an important role in highlighting data points which appear to be out-of-line with general country-level patterns and may warrant further investigation. These data may correctly reflect the effect of policy interventions or changes in societal conditions, but in many cases they have proved to be the result of issues with recording or classification and were subsequently corrected in the database.

Chapter 4

Delayed Reporting of Counts

This chapter is largely based on Stoner and Economou (2019) which was under review at the time of thesis submission.

In many fields and applications count data can be subject to delayed reporting. This is where the total count, such as the number of disease cases contracted in a given week, may not be immediately available, instead arriving in parts over time. For short term decision making, the statistical challenge lies in predicting the total count based on any observed partial counts, along with a robust quantification of uncertainty.

In this chapter we discuss previous approaches to modelling delayed reporting and present a multivariate hierarchical framework where the count generating process and delay mechanism are modelled simultaneously. Unlike other approaches, the framework can also be easily adapted to allow for the presence of under-reporting in the final observed count. To compare our approach with existing frameworks, one of which we extend to potentially improve predictive performance, we present a case study of reported dengue fever cases in Rio de Janeiro. Based on both within-sample and out-of-sample posterior predictive model checking and arguments of interpretability, adaptability, and computational efficiency, we discuss the advantages and disadvantages of each modelling framework.

4.1 Introduction

In many fields and applications where count data are collected, a situation can arise where the available reported count is believed to be less than or equal to the true count. Delayed reporting in particular is where the total observable count, which may still be less than the true count, is only available after a certain amount of time. In some situations information will trickle in over time so that the current total count gets ever closer to the true count, before eventually reaching the final total observable count.

An example of this situation is the occurrence of dengue fever, a viral infection

spread by mosquitoes, in Rio de Janeiro. Imagining we are at the end of week t, due to delayed reporting we have only observed a portion of the total observable number of cases which were contracted this week. A week from now, at time $t+1$, a further portion will become available and so on, such that after a number of weeks the total number of observed cases we have observed from week t eventually reaches a final total. Figure 4.1 shows an instance of the data, where we are at the end of week $t = 114$. The grey portions of each bar represent the yet unknown cases as of week t. For instance, we can see that for dengue cases that occurred in week $t-1$ we only have two weeks worth of information because we only have information that arrived in weeks $t-1$ and t, while for cases occurring in week $t-2$ we have three weeks worth of information and so on.

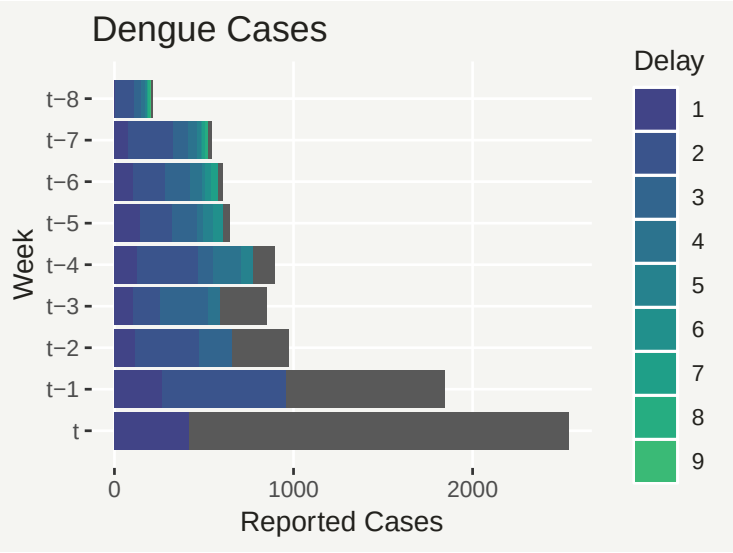

Figure 4.1: Bar plot of Rio de Janeiro dengue cases in the weeks leading up to time $t = 114$. The grey bars represent the total (as yet unobserved) number of reported cases, while the coloured bars show the number of cases reported in each week after the cases occurred.

Reporting delay becomes a problem when decisions need to be made based on the total count before it has been observed in its entirety. We can see in Figure 4.1, for example, that in the surveillance of dengue fever it can take months before the total observable number of cases contracted in a given week is known. This impedes the response time to severe outbreaks and puts lives at risk. It is therefore necessary to make predictions about the current state of the disease based on the partial number of cases observed (now-casting). This allows warnings to be issued and preparations to be made for predicted epidemics before they have been completely detected by the data. This motivates a statistical treatment of delayed reporting, with the goal of being able to predict the total count based on corresponding counts already observed. Further goals include predicting total counts which have not occurred yet (forecasting) and learning about the structure of the delay mechanism, so that improvements in reporting can be considered.

In this chapter we explore previous statistical approaches to modelling delayed reporting in count data, and discuss their strengths and weaknesses. We then propose a general framework for modelling count data with discrete-time delays, which is sufficiently flexible to be used for a range of data, including those with complex

spatio-temporal structures, and can be easily adapted to account for the presence of under-reporting in the final observed count. These approaches are assessed in the first instance through a simulation experiment, and then in a case study based on counts of dengue fever cases in Rio de Janeiro, Brazil. In both cases, we assess our proposed framework by means of posterior predictive checking, including of now-casting and forecasting performance in the case study, relative to existing approaches.

The chapter is structured as follows: Section 4.2 presents an overview of existing approaches to modelling count data suffering from discrete-time delayed reporting, in addition to a substantial extension to one of the existing approaches. In Section 4.3 we propose a general framework for modelling delayed reporting. Models representing these frameworks are applied to simulated data in Section 4.4, to assess their performance in an idealised scenario, while Section 4.5 presents their application to real dengue fever data from Rio de Janeiro. In Section 4.6, we discuss the potential issue of under-reporting in the final observed count and how the general framework from Section 4.3 can be adapted to account for it. Finally, Section 4.7 concludes with a discussion of interpretability, adaptability and ease of implementation.

4.2 Background

We begin by introducing some notation. Let $y_{t,s}$ be the total observable count occurring at temporal unit $t \in T$ and spatial unit $s \in S$. We refer to $y_{t,s}$ as the total observable count because, in some cases, the final count we observe may still be an under-representation of the true count, an issue we return to in Section 4.6. Suppose that after some (temporal) delay unit (e.g. one week) a portion of $y_{t,s}$ has been reported. We denote this first portion $z_{t,s,1}$. At the next delay unit we observe an additional portion of $y_{t,s}$, denoted as $z_{t,s,2}$. This continues such that at each delay unit $d \in \{1, ..., D\}$ we observe a count $z_{t,s,d}$, meaning the sum of the observed $z_{t,s,d}$ gets closer to the total count $y_{t,s}$.

4.2.1 Multinomial mixture approach

A sensible approach for modelling delayed reporting involves the idea of jointly modelling $z_{t,s,d}|y_{t,s}$ at the same time as the totals $y_{t,s}$. Höhle and an der Heiden (2014) propose modelling the delayed counts $\boldsymbol{z}_{t,s}|y_{t,s}$ as arising from a conditional Multinomial($\boldsymbol{p}_{t,s}, y_{t,s}$) distribution. Here $p_{t,s,d}$ is the expected proportion of $y_{t,s}$ which will be reported at delay d and is modelled as arising from Generalized-Dirichlet($\boldsymbol{\alpha}, \boldsymbol{\beta}$) (GD) distribution (Wong, 1998) where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are constant in time. The total observable count is also modelled explicitly as a latent Poisson variable in

the Multinomial model:

$$
\begin{aligned}
y_{t,s} \mid \lambda_{t,s} &\sim \text{Poisson}(\lambda_{t,s}) && (4.1)\\
\boldsymbol{z}_{t,s} \mid \boldsymbol{p}_{t,s}, y_{t,s} &\sim \text{Multinomial}(\boldsymbol{p}_{t,s}, y_{t,s}) && (4.2)\\
\boldsymbol{p}_{t,s} \mid \boldsymbol{\alpha}, \boldsymbol{\beta} &\sim \text{Generalized-Dirichlet}(\boldsymbol{\alpha}, \boldsymbol{\beta}) && (4.3)
\end{aligned}
$$

Wang et al. (2018) also apply this approach to the monitoring of foodborne diseases, while a similar approach (without the General-Dirichlet layer) can be found in Salmon et al. (2015).

However, the assumption that the Generalized-Dirichlet distribution is time-invariant can be viewed as a restriction in capturing any delay mechanism which varies systematically over time, potentially inhibiting nowcasting and forecasting precision. Höhle and an der Heiden (2014) seek to address this by presenting a second approach in which the Generalized-Dirichlet model is replaced with a more conventional logistic regression on the Multinomial probabilities:

$$
\begin{aligned}
\log\left(\frac{\nu_{t,s,d}}{1-\nu_{t,s,d}}\right) &= g(t,s,d) && (4.4)\\
p_{t,s,d} &= \nu_{t,s,d}\left(1-\sum_{i=1}^{d-1} p_{t,s,i}\right) && (4.5)
\end{aligned}
$$

where $g(t,s,d)$ is a linear combination of covariate effects. However, whilst this does allow the model to better capture heterogeneity in the delay mechanism over time, it is in part more restrictive. This is because in some applications the Multinomial delay model may be over-dispersed. We will discuss this issue in more detail in Section 4.3, where we propose a general framework which retains both the flexibility to capture spatio-temporal heterogeneity as well as the ability to appropriately separate variability in the delay mechanism from the model of the total count.

4.2.2 Conditional independence approach

An alternative approach, often used in the field of actuarial statistics for projecting ultimate losses from delayed insurance claims, is the Chain-Ladder method (Mack, 1993). The method is popular because it is easy to understand and is based entirely on empirical calculations. Renshaw and Verrall (1998) showed that the Chain-Ladder method can be presented as a Generalized Linear Model (Dobson and Barnett, 2018) of the following form:

$$
\begin{aligned}
z_{t,d} &\sim \text{Poisson}(\mu_{t,d}) && (4.6)\\
\log(\mu_{t,d}) &= \iota + \alpha_t + \beta_d && (4.7)
\end{aligned}
$$

This has been extended in various ways, for example to include potential covariates (see for instance England and Verrall (2002) and Barbosa and Struchiner (2002)).

These approaches however, are restrictive in the sense that they assume the delay structure is homogeneous in time. In reality, the way in which reporting is delayed, for example the proportion of cases reported in the first week, changes over time. The baseline Chain-Ladder model has therefore been extended to accommodate such non-homogeneities as well as spatial variability.

A highly flexible approach that in some sense generalises the Chain-Ladder, is the conditional independence approach where the partial counts $z_{t,s,d}$ ($d \in \{1, ..., D\}$) are modelled as independent quantities, conditional on any spatio-temporal or delay structures in their expected value. We refer to this as the Generalized Linear Model (GLM) approach, as it is effectively (conditional on dispersion parameters and random effects) a Negative-Binomial GLM (Dobson and Barnett, 2018) for the partial counts $z_{t,s,d}$:

$$z_{t,s,d} \mid \mu_{t,s,d}, \theta \sim \text{Neg-Bin}(\mu_{t,s,d}, \theta) \tag{4.8}$$

$$\log(\mu_{t,s,d}) = f(t,s) + g(t,s,d) \tag{4.9}$$

Here $f(t,s)$ and $g(t,s,d)$ can be linear (or indeed non-linear) combinations of covariate effects or random effects, including complex temporal, spatial and spatio-temporal structures. The former is intended to capture variation in the total counts $y_{t,s}$, while the latter is intended to capture variation in the delay mechanism. Aside from the flexibility of incorporating complex structures in the model for $\mu_{t,s,d}$, the key advantage of this approach is that it can be very easily implemented in a variety of frequentist frameworks (such as Generalized Additive Models, Wood (2017)), as well as Bayesian ones (such as Integrated Nested Laplacian Approximations (INLA) (Lindgren and Rue, 2015) and Markov Chain Monte Carlo (MCMC)). For example, Bastos et al. (2017) presents the application of this framework to dengue fever in Rio de Janeiro and to spatio-temporal severe acute respiratory infection (SARI) data in the state of Paraná (Brazil). Both were implemented in the Bayesian framework using INLA and in this case the framework was demonstrated to be a powerful tool for now-casting.

However, as $y_{t,s}$ is not modelled directly, inference is based on $y_{t,s} = \sum_{d=1}^{D} z_{t,s,d}$. Firstly, this means that uncertainty associated with the delay component of the GLM is potentially transferred through the summation of the $z_{t,s,d}$ into the uncertainty of the $y_{t,s}$. A consequence of this uncertainty propagation is that models such as (4.8)-(4.9) may result in forecasting uncertainty (for example as quantified by 95% prediction intervals) that is prohibitively large, particularly when forecasting into the future where no $z_{t,s,d}$ are available.

Furthermore, to obtain reliable inference for $y_{t,s}$, we would expect the model to capture $\text{Var}(y_{t,s})$ well. As $y_{t,s}$ is not modelled directly this is given by:

$$\text{Var}[y_{t,s}] = \text{Var}\left[\sum_{d=1}^{D} z_{t,s,d}\right] = \sum_{i=1}^{D}\sum_{j=1}^{D} \text{Cov}[z_{t,s,i}, z_{t,s,j}] \tag{4.10}$$

This means that capturing the variance of $y_{t,s}$ well relies on modelling the covariances of the $z_{t,s,d}$ well. The issue with this is that the covariances of the $z_{t,s,d}$ are restricted by the assumption that $z_{t,s,d}$ are independent, conditional on $\mu_{t,s,d}$. In many cases this may not be a valid assumption and consequently any inference based on $y_{t,s}$ is fundamentally flawed. Illustrating this problem and its potential consequences, we apply models representing the GLM, as well as a model representing the framework we propose in Section 4.3, to simulated data in Section 4.4. To potentially address this issue, in the following subsection we present an extension to the conditional independence approach, which may capture better the dependency structure of $z_{t,s,d}$ over d.

4.2.3 Extension of the conditional independence approach

We begin by noting that modelling $z_{t,s,d}$ with a Negative-Binomial distribution is equivalent to modelling $z_{t,s,d}$ as an over-dispersed Poisson quantity:

$$z_{t,s,d} \mid \mu_{t,s,d} \sim \text{Poisson}(\mu_{t,s,d}) \tag{4.11}$$

$$\mu_{t,s,d} \sim \text{Gamma}(\alpha_{t,s,d}, \beta_{t,s,d}) \tag{4.12}$$

In this form we can consider the variance of $z_{t,s,d}$ as the sum of the variance of the Poisson component and the variance of the Gamma component. A Gamma component which contributes more to the total variance corresponds to a lower value for the Negative-Binomial shape parameter and vice-versa. In the GLM framework we assume that both the Poisson and Gamma quantities are conditionally-independent across the delay indices d. Noting that in Bayesian hierarchical modelling the Gamma component is often approximated by a Log-Normal component, where the mean at the log-level includes an identically distributed Normal random effect, one approach to modelling conditional covariance between multivariate counts is to model the Poisson mean $\boldsymbol{\mu}_{t,s}$ as a Multivariate-Log-Normal quantity (Aitchison and Ho, 1989):

$$z_{t,s,d} \mid \mu_{t,s,d} \sim \text{Poisson}(\mu_{t,s,d}) \tag{4.13}$$

$$\log(\boldsymbol{\mu}_{t,s}) \sim \text{Multivariate-Normal}(\boldsymbol{\nu}_{t,s}, \boldsymbol{\Sigma}_{t,s}) \tag{4.14}$$

$$\nu_{t,s,d} = f(t,s) + g(t,s,d) \tag{4.15}$$

In this framework, which we refer to as the "GLM+ framework", the partial counts $z_{t,s,d}$ are still independent given $\mu_{t,s,d}$. However, at least some of the total covariance can be described explicitly by the Multivariate-Normal covariance structure. The implication of this is that the model may be better able to capture the covariance structure of the $z_{t,s,d}$, and consequently the variance of the total counts $y_{t,s}$, compared to the GLM framework.

In the following section, we present a general framework based on the Multinomial mixture approach, which retains the desirable merits of jointly modelling $z_{t,d,s}$

as well as the necessary flexibility to capture variability in the spatio-temporal and delay structures.

4.3 Generalized-Dirichlet-Multinomial Framework

Recall that $y_{t,s}$ denotes the true count occurring at temporal unit $t \in T$ and in spatial unit $s \in S$ and that $z_{t,s,d}$ denotes the observed count corresponding to $y_{t,s}$ with delay $d \in \{1, ..., D\}$. We begin by defining a Negative-Binomial model for the true counts:

$$y_{t,s} \mid \lambda_{t,s}, \theta_{t,s} \sim \text{Negative-Binomial}(\lambda_{t,s}, \theta_{t,s}) \tag{4.16}$$

$$\log(\lambda_{t,s}) = f(t,s) \tag{4.17}$$

with $f(t,s)$ the same as in Section 4.2. Modelling $y_{t,s}$ directly (as opposed to indirectly using the GLM), reduces the risk that $\text{Var}(y_{t,s})$ will not be captured well (as we will demonstrate in Section 4.4). However in order to make predictions about $y_{t,s}$ which have not yet been fully observed, we also need a model for the delayed counts $\boldsymbol{z}_{t,s}$ (which should provide partial information on the unobserved $y_{t,s}$):

$$\boldsymbol{z}_{t,s} \mid \boldsymbol{p}_{t,s}, y_{t,s} \sim \text{Multinomial}(\boldsymbol{p}_{t,s}, y_{t,s}). \tag{4.18}$$

Unlike the GLM approach, modelling the $\boldsymbol{z}_{t,s}$ in this way implies they are not assumed to be conditionally independent. In the simplest formulation of this framework, the $\boldsymbol{p}_{t,s}$ are not random but fixed, given any spatio-temporal structures or relevant covariates. However, this carries the risk of falsely confounding variability in the delay mechanism with variability in the true count $y_{t,s}$ when now-casting. We illustrate this by considering that the predictive distribution for unobserved totals $y_{t,s}$, conditional on partial counts $\boldsymbol{z}_{t,s}$, is given by:

$$p(y_{t,s}|\boldsymbol{z}_{t,s}) \propto p(\boldsymbol{z}_{t,s}|y_{t,s})p(y_{t,s}) \tag{4.19}$$

The issue is that $p(\boldsymbol{z}_{t,s}|y_{t,s})$ is Multinomial, which lacks flexibility in the variance as, conditional on $y_{t,s}$, both the mean and variance are defined solely by $\boldsymbol{p}_{t,s}$. As such, if there is excess variability (over-dispersion) in $\boldsymbol{z}_{t,s}|y_{t,s}$, this is likely to be erroneously absorbed by $p(y_{t,s})$. For example, if $z_{t,s,1}/y_{t,s}$ is too high for the Multinomial to reasonably capture given $p_{t,s,1}$, predictions of $y_{t,s}$ may be too high when now-casting. Moreover, as both the mean and correlation structure of $\boldsymbol{z}_{t,s}|y_{t,s}$ are exclusively defined by fixed $\boldsymbol{p}_{t,s}$, there is limited flexibility in capturing unusual covariance structures.

Both of these issues can be addressed by modelling $\boldsymbol{p}_{t,s}$ as a Generalized-Dirichlet($\boldsymbol{\alpha}_{t,s}, \boldsymbol{\beta}_{t,s}$) distribution, the probability density function of which is:

$$p(p_1, p_2, ..., p_k \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p_k^{\beta_{k-1}-1} \prod_{i=1}^{k-1} \left[\frac{p_i^{\alpha_i - 1}}{B(\alpha_i, \beta_i)} \left(\sum_{j=i}^{k} p_j \right)^{\beta_{i-1}-(\alpha_i+\beta_i)} \right]. \tag{4.20}$$

The resulting marginal model can be obtained by integrating out $\boldsymbol{p}_{t,s}$ to obtain a Generalized-Dirichlet-Multinomial or GDM($\boldsymbol{\alpha}_{t,s}, \boldsymbol{\beta}_{t,s}, y_{t,s}$) mixture distribution for $\boldsymbol{z}_{t,s}|y_{t,s}$, with probability mass function:

$$p(z_1, z_2, ..., z_k \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, y) = \frac{\Gamma(y+1)}{\Gamma(z_k+1)} \prod_{i=1}^{k-1} \left[\frac{\Gamma(z_i+\alpha_i)\Gamma(\sum_{j=i+1}^{k} z_j + \beta_i)}{B(\alpha_i, \beta_i)\Gamma(z_i+1)\Gamma(\alpha_i+\beta_i+\sum_{j=i}^{k} z_j)} \right]. \tag{4.21}$$

To be useful as a tool for nowcasting and forecasting, the model needs to be able to provide inference for $y_{t,s}$ conditional on any corresponding $z_{t,s,d}$ which have been observed (as well as any preceding $y_{t,s}$ which have been observed by the time of prediction). In a Markov Chain Monte Carlo implementation framework (such as the one used here) this is possible by sampling the corresponding $z_{t,s,d}$ which have not yet been observed as well as the unobserved $y_{t,s}$. However, to do the former we need to be able to sample from the conditional distributions $z_{t,s,d} \mid z_{t,s,1}, ..., z_{t,s,d-1}, y_{t,s}$. Fortunately, we can do this by defining and implementing the model in terms of the conditional structure of the GDM:

$$z_i \mid \boldsymbol{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, y \;\sim\; \text{Beta-Binomial}(\alpha_i, \beta_i, n_i = y - \sum_{j<i} z_j) \tag{4.22}$$

$$p(z_i \mid \boldsymbol{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, y) \;=\; \binom{n_i}{z_i} \frac{B(z_i+\alpha_i, n_i - z_i + \beta_i)}{B(\alpha_i, \beta_i)}. \tag{4.23}$$

To model structured variability in the delay mechanism, it makes sense to re-parametrise the Beta-Binomial in terms of its mean $\nu_{t,s,d}$ and dispersion parameter $\phi_{t,s,d}$, which relate to the parameters of the GDM by:

$$\alpha_{t,s,d} = \nu_{t,s,d}\phi_{t,s,d}; \qquad \beta_{t,s,d} = (1-\nu_{t,s,d})\phi_{t,s,d} \tag{4.24}$$

The intuition behind this characterisation is that, having already observed some delayed counts $z_{t,s,1}, ..., z_{t,s,d-1}$ corresponding to the true count $y_{t,s}$, then $\nu_{t,s,d}$ represents the proportion of the remaining (so far unreported) counts we expect to be reported in the next delay step d. Variability about $\nu_{t,s,d}$ is controlled by the dispersion parameter $\phi_{t,s,d}$. Both the mean and dispersion parameters can be generally characterised as functions of space, time and delay:

$$\log\left(\frac{\nu_{t,s,d}}{1-\nu_{t,s,d}}\right) \;=\; g(t,s,d) \tag{4.25}$$

$$\log(\phi_{t,s,d}) \;=\; h(t,s,d). \tag{4.26}$$

We can represent this approach in terms of the modular framework discussed in Section 1.3 as:

$$Y(\lambda_{t,s}, \theta_{t,s}) \rightarrow y_{t,s} \rightarrow Z(\boldsymbol{\nu}_{t,s}, \boldsymbol{\phi}_{t,s}) \rightarrow \boldsymbol{z}_{t,s} \tag{4.27}$$

where we have a latent model $Y(\lambda_{t,s})$ for the total observable counts $y_{t,s}$, with a further module $Z(\boldsymbol{\nu}_{t,s}, \boldsymbol{\phi}_{t,s})$ to take into account the delayed reporting mechanism.

In contrast to the GLM approach, predictive inference for the unobserved $y_{t,s}$ is based on both the delayed counts $\boldsymbol{z}_{t,s}$ and previous observed values $y_{t',s}$ for $t' < t$. In practice, using MCMC for model inference automatically generates predictive samples of the unobserved $y_{t,s}$ from $y_{t,s}|\boldsymbol{z}_{t,s}, y_{t',s}$. Furthermore, the delay mechanism does not appear in the model for $y_{t,s}$, meaning that associated variability does not propagate into the predictive inference for unobserved $y_{t,s}$.

4.4 Simulation Experiment

To illustrate the advantage of directly modelling the total recorded counts (to compare the GDM and GLM approaches), we apply four competing models to simulated data and assess their performance. The data was simulated from the following model:

$$\begin{aligned} y_i &\sim \text{Negative-Binomial}(\lambda = 100, \theta = 10) && (4.28)\\ \boldsymbol{z}_i \mid \boldsymbol{\pi}_i, y_i &\sim \text{Multinomial}(\boldsymbol{\pi}_i, y_i) && (4.29)\\ \boldsymbol{\pi_i} &\sim \text{Dirichlet}(\boldsymbol{\nu}\phi) && (4.30) \end{aligned}$$

In this model, the total counts y_i $(i = 1, ..., n = 100)$ arise from a Negative-Binomial model, with considerable over-dispersion compared to the Poisson distribution caused by the relatively low value for θ. These are all split into three partial counts $z_{i,j}$ $(j \in \{1, 2, 3\})$. These partial accounts arise from a Dirichlet-Multinomial mixture, with mean proportions $\boldsymbol{\nu} = (0.5, 0.2, 0.3)$ and a relatively low value for the dispersion parameter $\phi = 10$, such that the delay mechanism is also considerably over-dispersed compared to the Multinomial.

4.4.1 Competing models

Model 1 is a Negative-Binomial model for the total counts, with no model for the partial counts $z_{i,j}$. This is the baseline to which we will compare the others.

$$y_i \sim \text{Negative-Binomial}(\lambda, \theta) \tag{4.31}$$

Model 2 is a marginal Negative-Binomial model for the partial counts $z_{t,d}$, which ignores both the dependence between the $\boldsymbol{z}_t$ and the over-dispersion of the delay mechanism. This model is effectively the GLM approach described in Section 4.2.2, where the Negative-Binomial means are characterised as a product of marginal proportions ν_j and the total count rate λ.

$$z_{i,j} \sim \text{Negative-Binomial}(\nu_j\lambda, \theta) \tag{4.32}$$

Model 3 extends Model 2 by incorporating a Dirichlet model for the Multinomial proportions $\boldsymbol{\pi}_i$. The motivation behind this is that the Dirichlet can capture the

over-dispersion of the delay mechanism, as discussed in Section 4.3. This model is presented as something in-between the GDM and GLM approaches.

$$z_{i,j} \mid \pi_{i,j} \sim \text{Negative-Binomial}(\pi_{i,j}\lambda, \theta) \tag{4.33}$$

$$\boldsymbol{\pi}_i \sim \text{Dirichlet}(\boldsymbol{\nu}\phi) \tag{4.34}$$

Finally, **Model 4** is a Multinomial mixture (comparable to the GDM approach presented in Section 4.3), which accounts for over-dispersion with a Dirichlet model for $\boldsymbol{\pi}_i$.

$$y_i \sim \text{Negative-Binomial}(\lambda, \theta) \tag{4.35}$$

$$\boldsymbol{z}_i \mid \boldsymbol{\pi}_i, y_i \sim \text{Multinomial}(\boldsymbol{\pi}_i, y_i) \tag{4.36}$$

$$\boldsymbol{\pi_i} \sim \text{Dirichlet}(\boldsymbol{\nu}\phi) \tag{4.37}$$

For parameters λ, θ, and ϕ, non-informative Exponential prior distributions with mean 10000 were specified. Similarly, a uniform Dirichlet($\mathbf{1}$) prior was specified for $\boldsymbol{\nu}$. The models were implemented using NIMBLE and four chains were run for a total 20k iterations, discarding 10k as burn in. To give all models the best chance of capturing the correct parameter values, all chains were initialised at the true values. Convergence was assessed by computing the Multivariate Potential Scale Reduction Factor (described in more detail in Section 4.5.1) for all parameters and obtaining a value of less than 1.05 for each model.

4.4.2 Results

Figure 4.2 shows the posterior distributions for λ, θ and, where applicable ϕ, from each model. The dotted line shows the baseline Model 1.

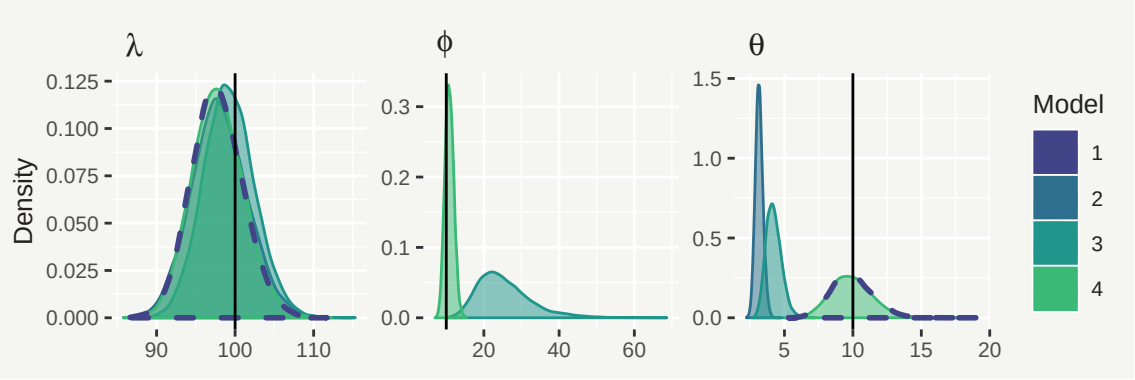

Figure 4.2: Posterior density plots of parameters λ (left), ϕ (centre) and θ (right), from each model. The dotted line shows the baseline Model 1.

Whilst all models were able to correctly capture the true value of λ, only Models 1 (baseline) and 4 were able to correctly capture the Negative-Binomial dispersion parameter θ. In the case of Model 2 (representative of the GLM approach), this is likely because the lack of an over-dispersion model in the delay mechanism meant θ

had to absorb the additional variability. Model 3 was closer to the true value of θ but still very far off (despite its Dirichlet model for over-dispersion). The consequence, as shown in Figure 4.3 is that both Models 2 and 3 grossly over-estimate the variance of the total counts y_t, when simulating posterior replicates.

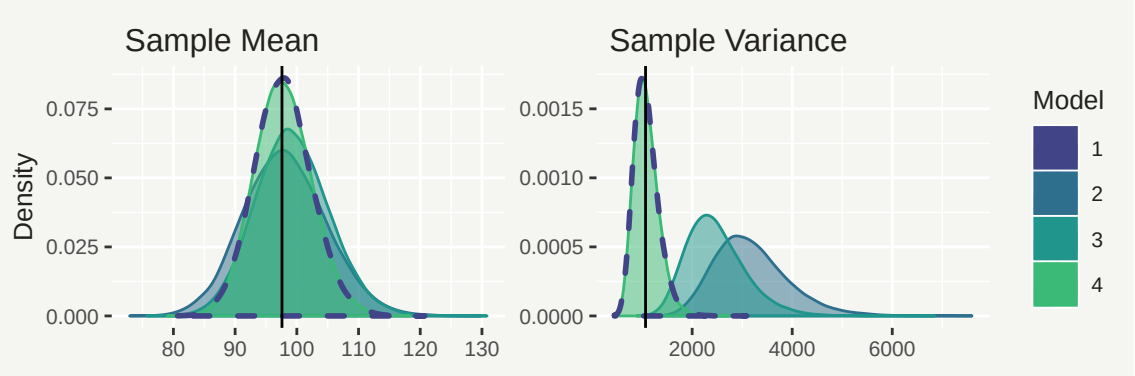

Figure 4.3: Density plots of the sample mean (left) and variance (right) of posterior replicates of total counts y_i, from each model. The dotted line shows the baseline Model 1.

In contrast, by both allowing for an over-dispersed delay mechanism and directly modelling the total counts y_t, Model 4 (representative of the GDM approach) is able to match the baseline model in capturing θ correctly and, as a result, the variance of the simulated y_t.

The conclusion we draw from this experiment is that, in a situation where the Multinomial model for the partial (delayed) counts $z_{t,d}$ is over-dispersed, failing to take into account this over-dispersion and/or ignoring the dependency structure of the $\boldsymbol{z}_t$ (by modelling them as conditionally independent, as in the GLM approach) can translate to substantial over-estimation of the variance of the total counts y_t.

In the subsequent section we will apply equivalent GDM, GLM and GLM+ models to dengue fever data, discussing their relative merits with respect to performance in model checking, now-casting and forecasting.

4.5 Case Study

Dengue fever is a viral infection, transmitted by mosquitoes, which has flu-like symptoms that may evolve into a potentially fatal condition known as severe dengue (World Health Organization, 2018b). The disease causes a major burden for the population it affects, particularly in Brazil, which reports more dengue cases than any other country (Silva et al., 2016). Effective response to dengue requires early detection (World Health Organization, 2018b), so it is important that healthcare providers are able to prepare themselves for a possible outbreak. Though the reporting of dengue cases to the Brazilian national surveillance system (SINAN) is mandatory (Silva et al., 2016), it can take weeks or even months of delay for the

number of reported cases occurring in a given week to approach a final count. For this reason, statistical delayed-reporting models are used to correct delays and predict outbreaks before the total count is available (Bastos et al., 2017).

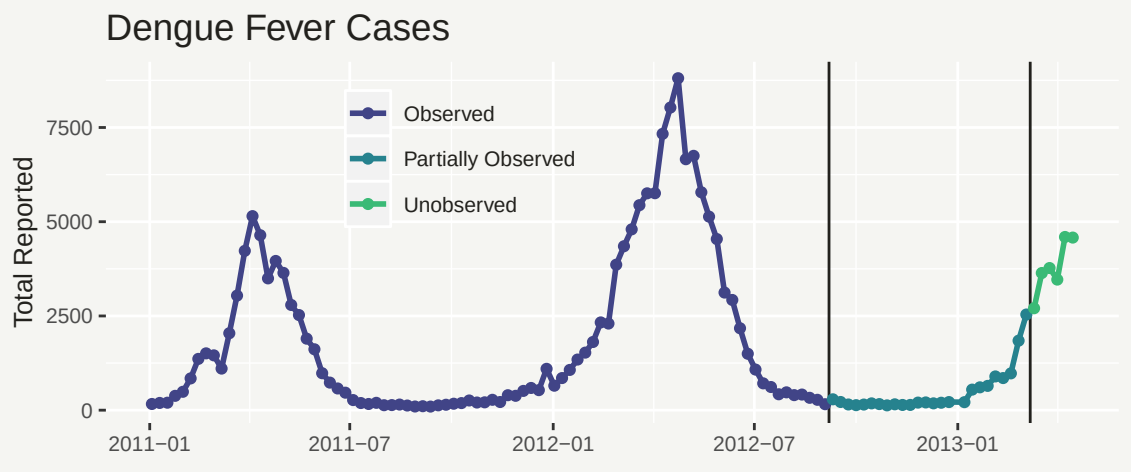

Figure 4.4: Total number of reported dengue cases from 2011 onwards in Rio de Janeiro. Different colours represent which data are fully observed, partially observed or unobserved at week $t = 114$ (March 2013).

Here we consider data on dengue fever cases in Rio de Janeiro, Brazil, occurring in weeks $t = 1$ (week commencing the 3rd of January 2011) to $t = 120$ (week commencing the 15th of April 2013). For illustration, we imagine that present day is week $t = 114$ (week commencing the 4th of March 2013). Furthermore, we consider the total observable count to be the number of cases observed after 6 months (26 weeks) worth of data (in addition to the number of cases reported in the week of occurrence). With present day being week $t = 114$, this implies we have 88 weeks of fully observed total counts y_t. Total counts occurring in weeks $t = 89$ to $t = 114$ are only partially observed and must be predicted based on the partial observations (now-casting). Total counts y_t after present day ($t = 114$) have not yet occurred and so they are completely unobserved. This is the forecasting period.

The time series of counts is illustrated in Figure 4.4, with different colours corresponding to the three different periods. There is some evidence of seasonality in the data, with outbreaks starting around the beginning of the calendar year and ending approximately 6 months later. This reflects the fact that the incidence of dengue fever is thought to depend heavily on the time of year and climatological conditions (Morales et al., 2016). We can also see some non-seasonal temporal structure, meanwhile, with the outbreak in 2012 being more severe than the one in 2011.

The top-left panel in Figure 4.5 shows the proportion of dengue cases reported in the week they occurred (first week) plotted against time, while the other panels show the proportion of cases reported a week after they occurred (second week), the following (third) week, and the week after that (fourth), respectively. We can see strong evidence of temporal structure in the delay mechanism, with the average

proportion reported in the first week steadily dropping throughout 2011, reaching its lowest point at the start of 2012 before beginning to rise again.

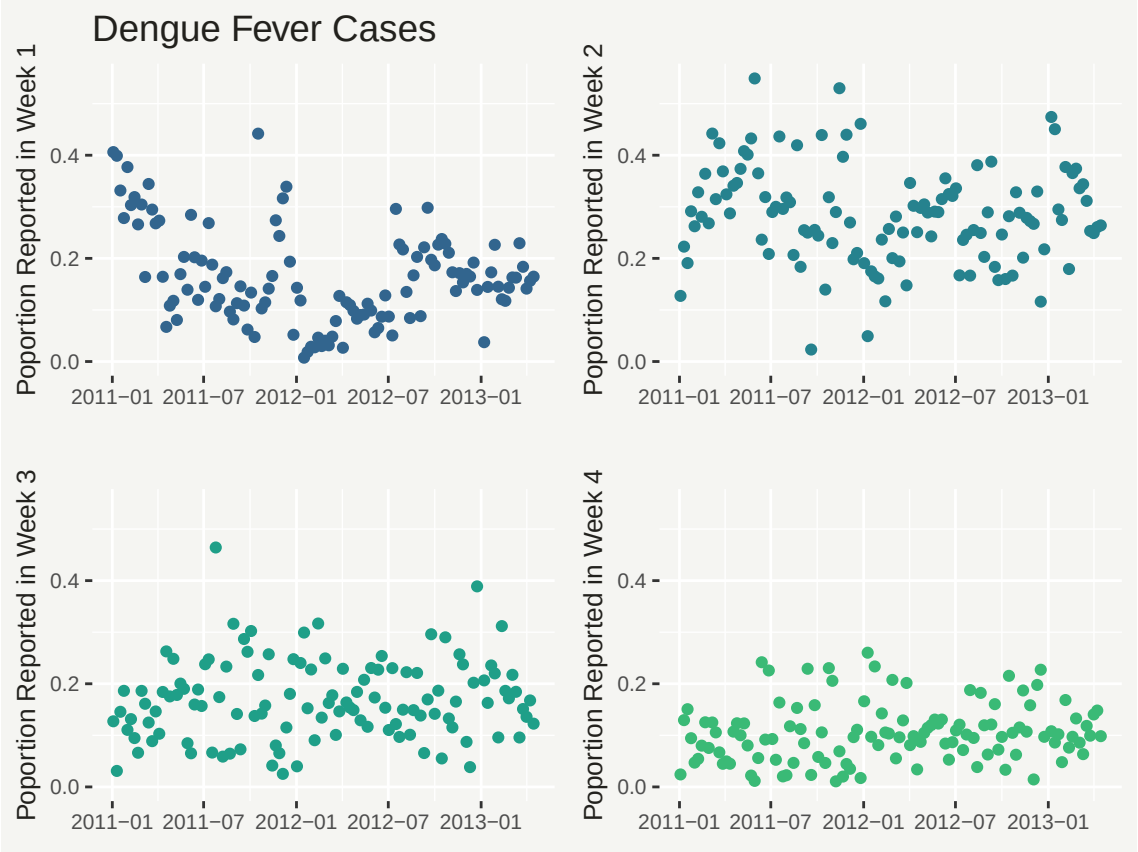

Figure 4.5: Proportion of dengue cases reported in the first week (the week in which they occurred, (top-left), the second week (top-right), the third week (bottom-left), and the fourth week (bottom-right).

4.5.1 Formulation of competing models

We now model this data using three comparable models (in terms of flexibility and interpretation), namely the GDM, GLM and GLM+. Modelling every partial count $z_{t,d}$ (in this case all 27 weeks) will result in the greatest predictive precision, though this comes at a high computational cost. Instead, if the total y_t is almost entirely observed after D delay steps, it may be more pragmatic to model only counts $z_{t,d}$ up to $d = D$ as well as the sum of the remaining counts $z_{t,D+1} = y_t - \sum_{d=1}^{D} z_{t,d}$. In the GDM approach this is achieved by only including the conditional models for the first D partial counts, such that the remainder is modelled implicitly, while in the GLM and GLM+ approaches this can be achieved by modelling $z_{t,D+1}$ in the same way as the individual counts.

One way to make this decision is to consider the proportions of each observed y_t reported after each delay step. Figure 4.6 shows the 20%, 40%, 60%, and 80% quantiles of the proportions of the total dengue cases reported after each delay step. By looking at the 20% quantiles of these proportions we can see that the vast

majority (over 80%) of total dengue cases are covered after $D = 8$ weeks worth of data 80% of the time, with little to be gained unless many more weeks are considered. For this reason we choose to model only the first 8 weeks individually.

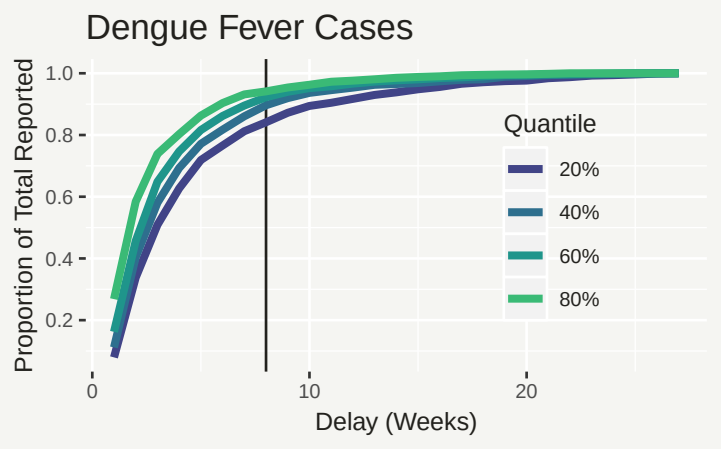

Figure 4.6: Quantiles of the proportions of fully observed ($t = 1, ..., 88$) total dengue cases y_t covered by $\sum_{d=1}^{D} z_{t,d}$ after each additional week of data.

The model based on the GDM framework is defined by:

$$\begin{aligned} y_t &\sim \text{Negative-Binomial}(\lambda_t, \theta) && (4.38)\\ \log(\lambda_t) &= \iota + \alpha_t + \eta_t && (4.39)\\ \boldsymbol{z}_t \mid y_t &\sim \text{GDM}(\boldsymbol{\nu}_t, \boldsymbol{\phi}, y_t) && (4.40)\\ \log\left(\frac{\nu_{t,d}}{1-\nu_{t,d}}\right) &= \psi_d + \beta_{t,d} && (4.41) \end{aligned}$$

Where $\boldsymbol{\nu}_t$ and $\boldsymbol{\phi}$ are the expectations and dispersions parameters of the Beta-Binomial conditional distributions, as described in (4.22)-(4.26).

The model based on the GLM framework is defined by:

$$\begin{aligned} z_{t,d} &\sim \text{Negative-Binomial}(\mu_{t,d}, \theta_d) && (4.42)\\ \log(\mu_{t,d}) &= \iota + \alpha_t + \eta_t + \psi_d + \beta_{t,d} && (4.43) \end{aligned}$$

The model based on the GLM+ framework is defined by:

$$\begin{aligned} z_{t,d} &\sim \text{Negative-Binomial}(\mu_{t,d}, \theta_d) && (4.44)\\ \log(\boldsymbol{\mu}_t) &\sim \text{Multivariate-Normal}(\boldsymbol{\nu_t}, \boldsymbol{\Sigma}) && (4.45)\\ \nu_{t,d} &= \iota + \alpha_t + \eta_t + \psi_d + \beta_{t,d} && (4.46) \end{aligned}$$

In all models η_t is a penalized cyclic cubic spline (Wood, 2017) defined over weeks $1, ..., 52$, which represents the effect of the time of year on the total number of reported dengue cases, and α_t is a penalized cubic spline defined over the whole temporal range. The latter is designed to capture non-seasonal temporal structure in the rate of total reported dengue cases and is constrained to be linear beyond the final knot so that it can be used for forecasting. The effects $\beta_{t,d}$ are defined by a different penalized cubic spline (each with its own smoothness penalty) for each delay index d, intended to capture temporal changes in the delay mechanism over time. As

discussed in Wood (2016), the coefficients of each spline are assigned a Multivariate-Normal prior distribution and are penalized to prevent excessive wiggliness through an unknown penalty parameter τ (the scaling factor of the Multivariate-Normal prior precision matrix). The re-parametrisation $\sigma = 1/\sqrt{\tau}$ is potentially more interpretable for the purpose of specifying a prior distribution, where smaller values of σ correspond to a stricter penalty on how flexible the smooth function is. The splines are centred to have zero-mean, and as such the models include the fixed effects ι and ψ_d as intercepts.

The Negative-Binomial dispersion parameters (θ_d and θ) were assigned relatively non-informative Exponential(0.01) prior distributions. The GDM dispersion parameters ϕ_d were assigned Log-Normal$(2, 2)$ prior distributions, such that most of the prior density is over values of ϕ_d which result in a modest contribution from the Generalized-Dirichlet component to the overall variance of the GDM, without ruling out higher values which correspond to a Multinomial situation. Relatively non-informative Normal$(0, 10^2)$ prior distributions were specified for the global intercept parameter ι and also for the delay-specific intercept parameters ψ_d. In the GDM model, the intercept parameters ψ_d represent the means of relative proportions at the logistic level. For these parameters we specified Normal prior distributions with the means chosen so that the prior mode implies approximately equal amount of cases being reported in each week of delay, with the variance chosen so that they are relatively non-informative. We specified Half-Normal$(0, 1)$ prior distributions for the penalty parameters of splines α_t and η_t. This imposes a relatively strong smoothness penalty on the effects α_t and η_t, which are supposed to capture medium-to-long term trends in the incidence of dengue cases. We relaxed this penalty slightly for the effects $\beta_{t,d}$ by specifying weaker Half-Normal$(0, \sqrt{2})$ priors. Finally, for the Multivariate-Normal covariance of $\log(\boldsymbol{\mu}_t)$ in the GLM+ model, we specified a fairly weak Inverse-Wishart prior with an identity scale matrix (dimension $D + 1$) and $D + 2$ degrees of freedom.

All code was written and executed using R (R Core Team (2018)) and all three models were implemented using NIMBLE (de Valpine et al., 2017), a facility for highly flexible implementation of MCMC. The model matrices for the splines were set up using the package `jagam` (Wood, 2016). Four MCMC chains were run from different initial values and with different random number generator seeds, until the following convergence criteria were met.

For each model, convergence of the four chains was assessed by visual inspection of trace plots and by computing the Multivariate Potential Scale Reduction Factor (MPSRF) (Brooks and Gelman, 1998) of a selection of model parameters. This is a scalar measure which generalizes the PSRF (detailed in Section 2.3.4) to sets of more than one model parameter, with the same interpretation that a value close to 1 indicates convergence. As with the PSRF, by convention a value of less than 1.05

for the MPSRF is assumed to represent convergence.

- For the GDM model, we computed the MPSRF of the set of parameters including every 10th λ_t ($\lambda_{10}, \lambda_{20}, ...$), θ, every 10th $\beta_{t,d}$ and the ϕ_d. The model was run for a total of 400k iterations, discarding the first 200k as burn-in and thinning by 20 to save memory. The MPSRF was computed to be 1.05 indicating that the model had converged.

- For the GLM model, we computed the MPSRF of the set including every 10th $\mu_{t,d}$ and the θ_d. The model was run for a total of 800k iterations, discarding the first 400k as burn-in and thinning by 40 to save memory. The MPSRF was computed to be 1.04.

- For the GLM+ model, we computed the MPSRF of the set including every 10th $\mu_{t,d}$. The model was run for a total of 800k iterations, discarding the first 400k as burn-in and thinning by 40 to save memory. The MPSRF was computed to be 1.02.

4.5.2 Results

To compare the models we will begin by exploring which aspects of the results are similar. Figure 4.7 shows the posterior mean predicted temporal effect (α_t) as well as the seasonal effect (η_t) from the GDM, GLM and GLM+ models, with associated 95% credible intervals, on the incidence rate dengue cases (at the log-scale). Both effects are very similar in shape between the three models: in the left panel we can see that all models suggest a persistent increase in dengue incidence in 2012, which makes sense given the more severe outbreak shown in Figure 4.4, while the right panel shows a strong seasonal effect in all models, with a much higher incidence rate in the first half of the year than the second. Interestingly the seasonal effect is less certain, though still strong, for the GDM model compared to the GLM and GLM+ models. Given that there are only approximately two years of fully observed data, the uncertainty in the GDM model's seasonal effect seems more reasonable.

Similarly, Figure 4.8 shows that, although not perfectly comparable because the models use different link functions (logistic for GDM and log for GLM and GLM+), the temporal effects on the number of cases reported in the first week are very similar between the three models. For example, all three models show a distinct drop in proportion of cases reported in the first week during the 2012 outbreak.

We now move on to ways in which the models differ. Recall from Section 4.2 that, in the GLM framework, capturing the distribution of the true counts $y_{t,s}$ well relies on a potentially restrictive assumption that the delayed counts $z_{t,s,d}$ are conditionally independent. In contrast, by modelling the total counts $y_{t,s}$ explicitly, the GDM framework has more flexibility to capture their distribution well. Similarly, the

addition of a covariance model in the GLM+ framework means that it may be able to capture the covariance of the partial counts $z_{t,d}$, and consequently the variance of the total counts, better than the GLM framework.

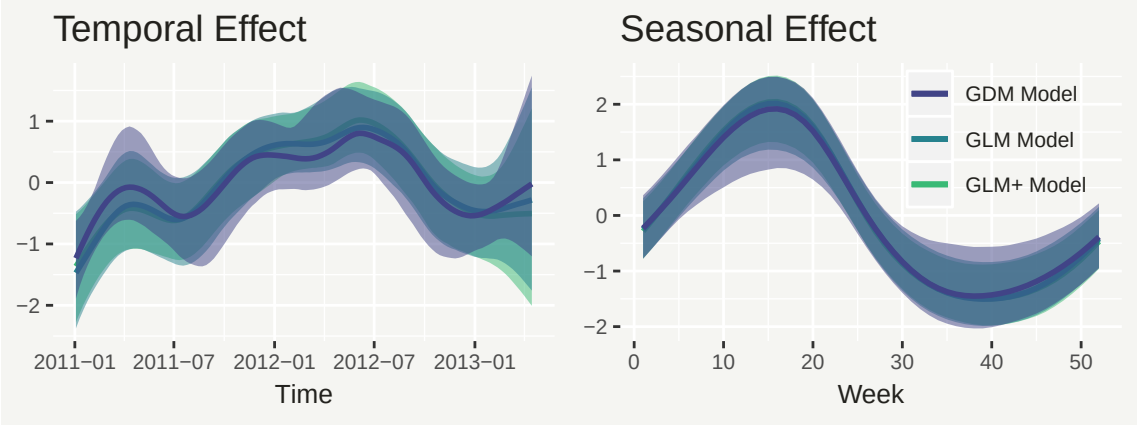

Figure 4.7: Posterior mean temporal (α_t) and seasonal (η_t) spline effects on the incidence rate of dengue cases, from the GDM, GLM and GLM+ models, with associated 95% credible intervals.

We use in-sample posterior predictive checking (Gelman et al., 2014) to assess the fit of the models to the data. This is done by simulating replicates of the observed partial counts $\tilde{z}_{t,d} \mid z_{t,d}$ and the fully observed (weeks 1-88) total dengue counts $\tilde{y}_t \mid y_t$ from the respective predictive distributions. We can then see if particular statistics of the observed data are captured well, by comparing them to the distribution obtained by computing the corresponding statistics of the replicates.

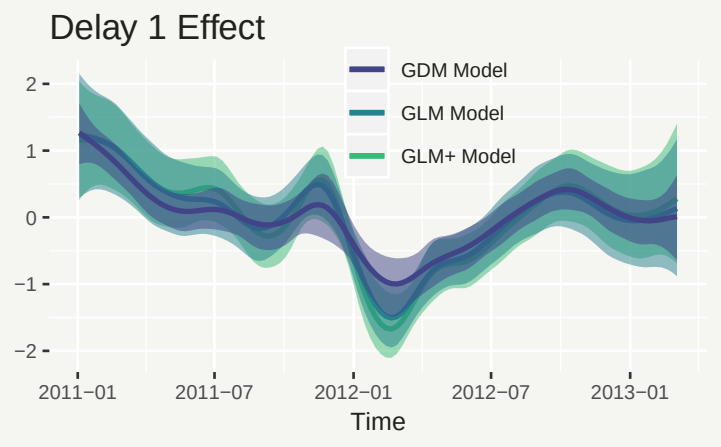

Figure 4.8: Posterior mean delay spline effect $\beta_{t,1}$ corresponding to counts reported in the first week $z_{t,1}$, from the GDM, GLM and GLM+ models, with associated 95% credible intervals.

We begin by looking at the covariance of the partial counts $z_{t,d}$ and the covariance of the proportion reported in each week $z_{t,d}/y_t$. For each sets of replicates, we compute the sample covariance of these two quantities, resulting in a distribution of samples for each individual covariance $\text{Cov}[\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{z}}_j]$ and $\text{Cov}[\tilde{\boldsymbol{z}}_i/\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}_j/\tilde{\boldsymbol{y}}]$. The left column of Figure 4.9 shows the mean difference between the replicate covariances and the observed covariances, while the right column shows the mean squared difference between the replicate covariances and the observed covariances. For both

the covariance of the partial counts and the covariance of the proportion reported in each week, we can see that the GDM model is the least biased (potentially even unbiased for the proportion reported in each week) and the most precise (lowest mean squared error).

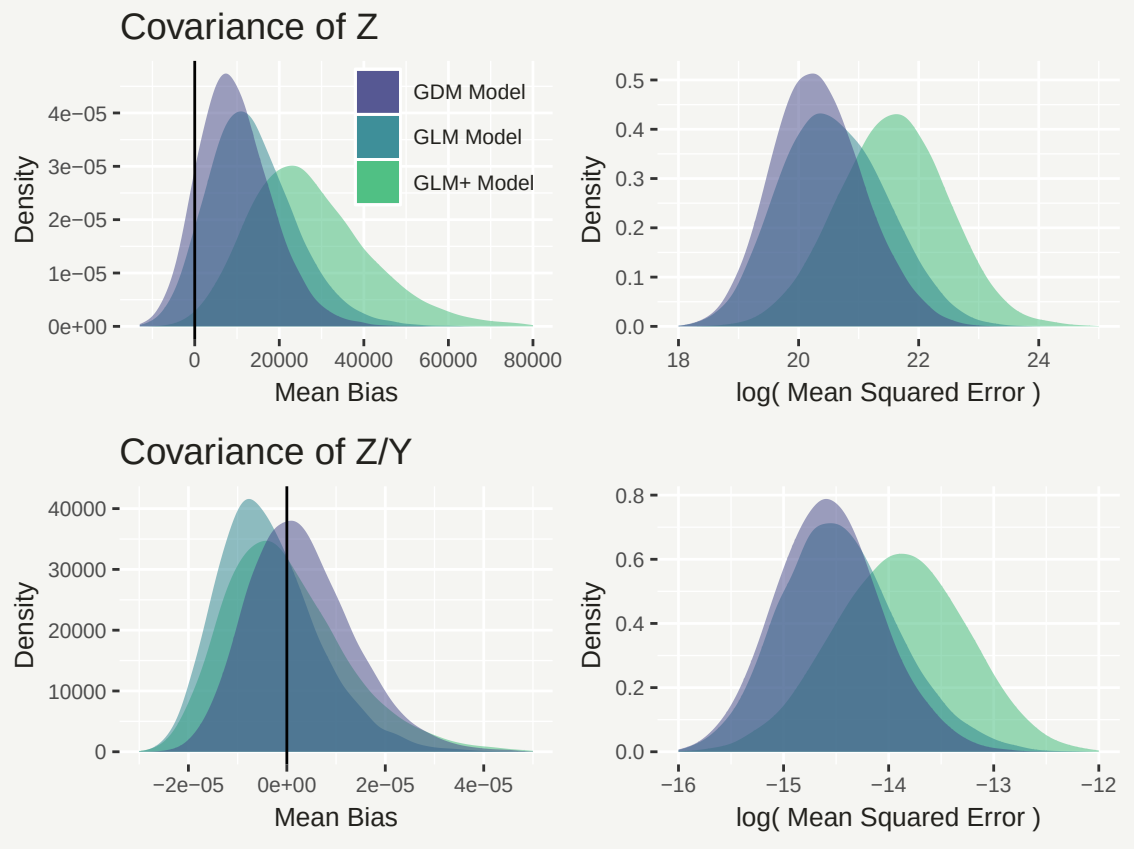

Figure 4.9: Density plots of the mean bias (left column) and the logarith of the mean squared error (right column) of the covariance of the partial counts $z_{t,d}$ and the proportion reported in each week $z_{t,d}/y_t$.

Similarly, the left and central panels of Figure 4.10 show density estimates of the distribution of the sample mean and the sample variance, respectively, of the replicate total counts $\tilde{y}_t$. We can see that in both cases the observed statistic, represented by a vertical line, is captured best by the GDM model, with the GLM faring better than the GLM+ model. This is a surprising result, given that the GLM+ has more flexibility than the GLM to capture the covariance structure of the partial counts $z_{t,d}$. The right panel of Figure 4.10 shows posterior means of the sorted replicates, with 95% prediction intervals. In this plot we can clearly see that, while the distribution of the total counts is captured best by the GDM and adequately well by the GLM, the GLM+ has an excessively heavy upper tail, compared to the data. This difference is likely because in the Poisson-Log-Normal mixture the logarithm of the Poisson mean is symmetric, compared to negatively skewed in the Poisson-Gamma mixture.

Recall that two important uses of delayed-reporting models are the prediction of

total counts $y_{t,s}$ which have occurred but haven't yet been fully observed (nowcasting) and the prediction of total counts which have not yet occurred (forecasting). In this case study we imagine we are in week 114 and we would like to predict the number of dengue cases in recent weeks (e.g. y_{114}) as well as to predict dengue cases over the next 6 weeks. Figure 4.11 shows the posterior median predicted number of dengue cases y_t from the three models, with associated 95% posterior predictive intervals. We can see that, whilst the median predictions from all three models are virtually identical, the model with the least predictive uncertainty, in both the now-casting range and forecasting range, is the GDM, making the GDM forecast potentially most useful to decision-makers. Notably, the GLM+ is far closer to the GDM in terms of certainty than the GLM, suggesting our extension may have improved now-casting and forecasting precision. However, we would consider the GDM's quantification of uncertainty most trustworthy, given its favourable results in the in-sample predictive checking.

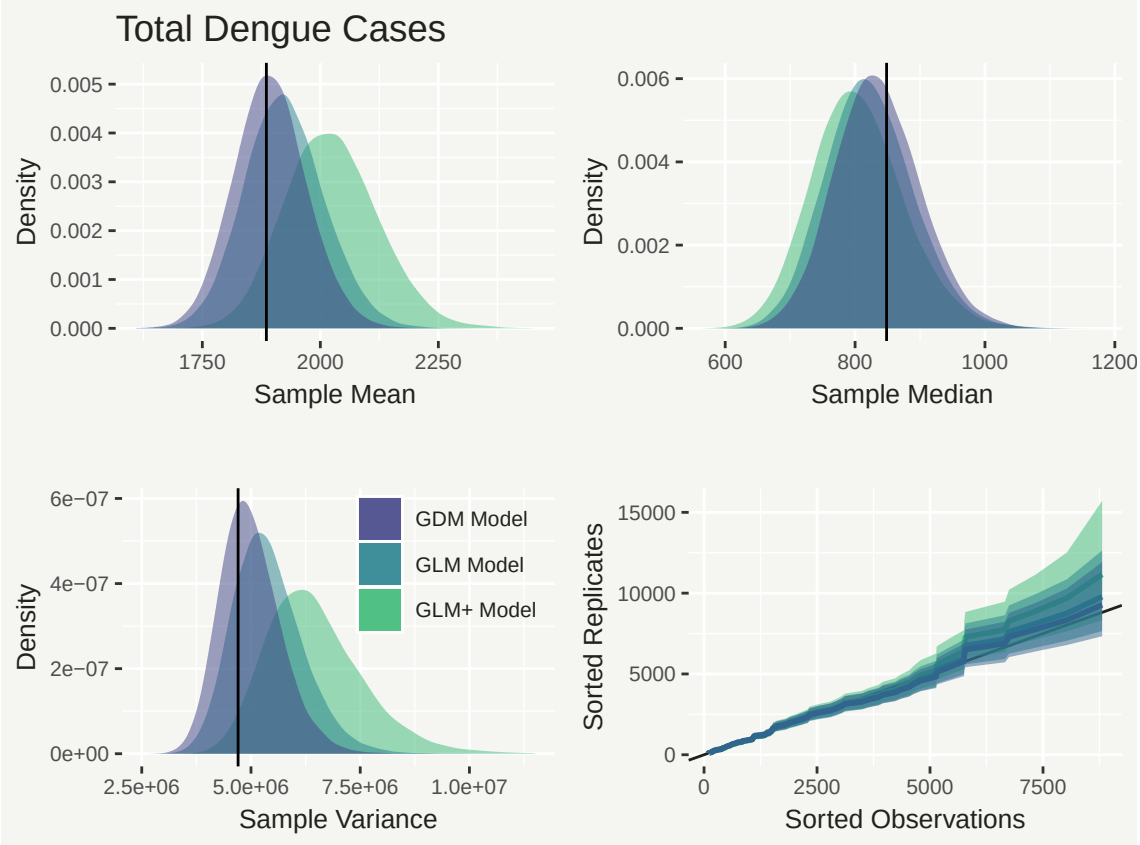

Figure 4.10: The left and central panels show density plots of the sample mean and sample variance of the posterior replicates of the fully observed (weeks 1-104) total dengue cases (y_t) from the GDM and GLM models. The vertical lines represent the corresponding statistics from the observed data. The right panel shows the mean of replicates of the total dengue cases y_t, from the GDM and GLM models, with associated 95% posterior predictive intervals.

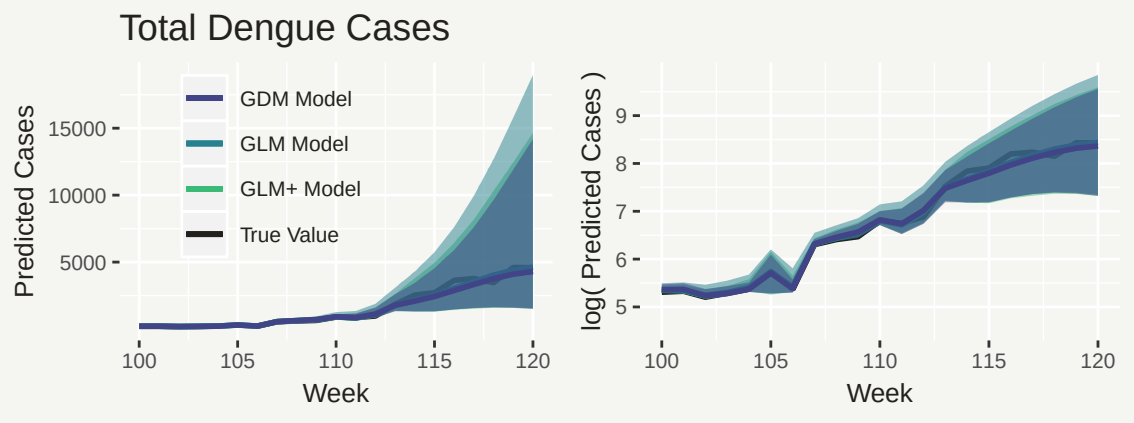

Figure 4.11: Posterior median predictions of the unobserved total dengue cases y_t, from the GDM, GLM and GLM+ models, with associated 95% posterior predictive intervals. Predictions beyond week $t = 114$ are forecasting without any observed partial counts $z_{t,d}$.

4.5.3 Comparison with other approaches

By this point we have demonstrated several ways, for this data, in which the GDM framework improves over the GLM framework our own extension of it, the GLM+ framework. It remains to show that the increased flexibility of the GDM over other approaches discussed in Secton 4.2 leads to tangible improvements in this example. Recall that one method presented by Höhle and an der Heiden (2014) and others, is to treat the parameters of the Generalized-Dirichlet component as stationary in time. As we saw in Figure 4.5, there is substantial variation over time in the proportion of dengue cases reported in the first week. This structure would not be captured by assuming time-stationarity in the Generalized-Dirichlet model, inevitably leading to poorer nowcasting and forecasting performance.

An alternative suggestion was to model the proportion of cases reported at each delay level in each week using a conventional Multinomial logistic regression, removing the additional variability provided by the Generalized-Dirichlet component.

One way to assess the contribution of the GD variance is to simulate posterior replicates of the proportion reported in each week of delay ($z_{t,d}/y_t$) both from the GDM using the posterior samples for the dispersion parameters ϕ_d and again from the same model but in the limiting case when $\phi_d \to \infty$, such that the joint conditional distribution of $z_{t,d}$ is Multinomial. Figure 4.12 shows 95% posterior predictive distributions for the proportion of dengue cases reported after 1 (top) and 2 (bottom) weeks of delay for both the model with GD variance and without. We can see that without the GD variance an excessively high number of points are not captured by the prediction intervals. Also shown are the 95% prediction interval coverages: the proportion of observations which lie within their corresponding 95% prediction intervals. The coverages with the GD variance are just over 95%, indicating a good fit to this data, while less than two-thirds of points are covered without the GD

variance.

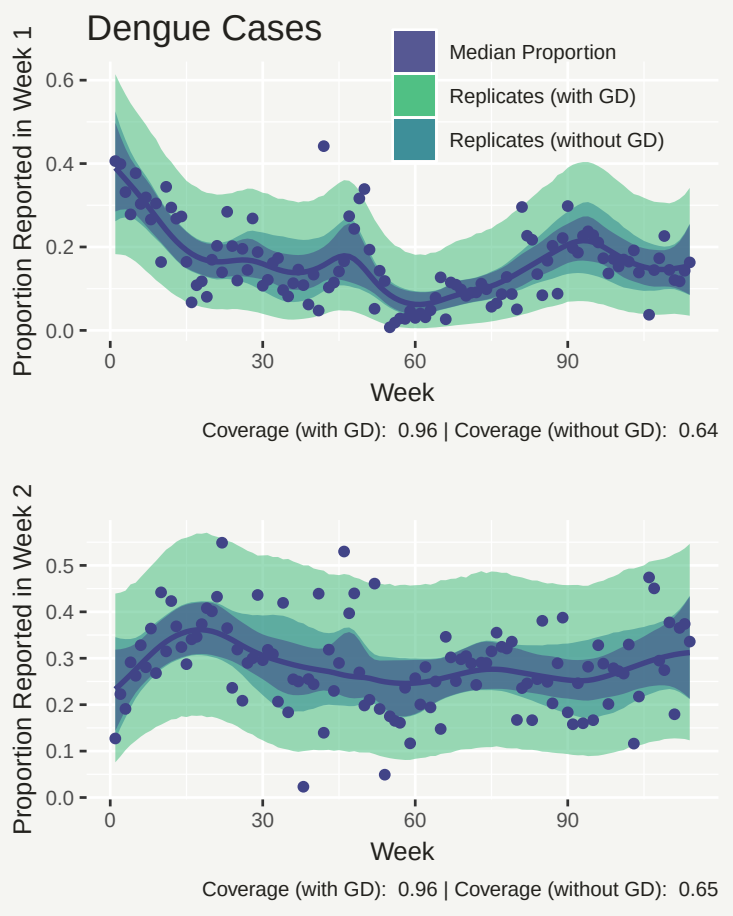

Figure 4.12: Posterior median proportion, from the GDM model, of dengue cases reported in the first (top) and second (bottom) weeks after incidence, with associated 95% credible intervals. Also shown are 95% posterior predictive intervals of the proportion reported in the first and second weeks from the GDM model with and without the additional variance from the Generalized-Dirichlet layer.

4.6 Under-reporting

An added challenge that occurs in data that are subject to reporting delay is that, in some situations, the final observed total count $y_{t,s}$ may still be a (substantial) under-estimate of the true count. In disease surveillance, this may translate to many cases never being reported, leading to a biased understanding (underestimation) of the actual magnitude of outbreaks. For instance, although reporting of dengue cases to the national surveillance system (SINAN) is mandatory, research suggests that the reported total may be substantially lower than the true number of dengue cases, owing to issues such as patients not seeking healthcare (Silva et al., 2016).

In our conceptual framework for taking into account flawed observation mechanisms, addressing this issue is just a case of incorporating an additional module to take into account the under-reporting:

$$X(\lambda_{t,s}, \theta_{t,s}) \to x_{t,s} \to Y(\pi_{t,s}) \to y_{t,s} \to Z(\boldsymbol{\nu}_{t,s}, \boldsymbol{\phi}_{t,s}) \to \boldsymbol{z}_{t,s} \tag{4.47}$$

Before, we assumed that the total observable counts arise from a Negative Binomial model. Now we assume this model ($X(\lambda_{t,s}, \theta_{t,s})$) instead generates unobserved true

counts $x_{t,s}$, such that $y_{t,s} \leq x_{t,s}$. We now need an under-reporting model ($Y(\pi_{t,s})$) to transform the true counts into the total observable counts.

Fortunately, we already have such a model: the hierarchical model discussed in Chapter 2. The complete model for delayed reporting and under-reporting is given by:

$$\begin{aligned} x_{t,s} \mid \lambda_{t,s}, \theta &\sim \text{Negative-Binomial}(\lambda_{t,s}, \theta) && (4.48)\\ y_{t,s} \mid x_{t,s}, \pi_{t,s} &\sim \text{Binomial}(\pi_{t,s}, x_{t,s}) && (4.49)\\ \log\left(\frac{\pi_{t,s}}{1-\pi_{t,s}}\right) &= i(t,s) && (4.50)\\ \boldsymbol{z}_{t,s}|y_{t,s} &\sim \text{GDM}(\boldsymbol{\nu}, \boldsymbol{\phi}, y_{t,s}) && (4.51) \end{aligned}$$

such that $\lambda_{t,s}$ represents the incidence rate of the true count $x_{t,s}$ (as opposed to the total observed count $y_{t,s}$) and $\pi_{t,s}$ represents the reporting rate. As illustrated in Section 2.2.2, both covariates and random effects can be used to model this reporting rate at the logistic level, represented by the generic function $i(t,s)$ in (4.50).

As discussed in Section 2.2.2, without any observations for the true count $x_{t,s}$ the model is not identifiable between a high reporting rate $\pi_{t,s}$ and a low incidence rate $\lambda_{t,s}$, or vice versa. However, this can be resolved through the use of at least one informative prior (such as for the overall reporting rate across the whole time series, as discussed in Stoner et al. (2019a)).

Using this approach means that policy and intervention can be based on predictions for the true number of cases, taking into account both the delayed reporting and under-reporting mechanisms, to reduce the risk of an undersized response. Note further, that allowing for under-reporting in the total count would be much less straightforward using the GLM and GLM+ approaches, mainly because the totals $y_{t,s}$ are not modelled explicitly.

4.6.1 Application to dengue

To highlight the ease of integrating our framework for under-reporting into our framework for delayed-reporting, we extend the GDM model for the dengue data to allow for hypothetical under-reporting.

Starting with the GDM model presented in Section 4.5.1, we reassign the Negative-Binomial model to the true (unobserved) total dengue counts x_t:

$$x_t \sim \text{Negative-Binomial}(\lambda_t, \theta) \quad (4.52)$$

For the total reported counts y_t, we then specify a Binomial model with reporting probability π:

$$y_t \mid x_t \sim \text{Binomial}(\pi, y_t) \quad (4.53)$$

If we had covariates which may relate to the under-reporting mechanism, we could incorporate these in a logistic model for π, as suggested by (4.50). These aren't available to us, so to simply illustrate the ease of combining the two flawed-observation modules, we model the reporting probability π as constant for all weeks. For this we specify a symmetrical Beta(100,100) prior, representing a hypothetical expectation for the reporting rate of 50%, with a standard deviation of approximately 3.5%. The remainder of the model is then the same as before:

$$\log(\lambda_t) = \iota + \alpha_t + \eta_t \tag{4.54}$$

$$\boldsymbol{z}_t \mid y_t \sim \text{GDM}(\boldsymbol{\nu}_t, \boldsymbol{\phi}, y_t) \tag{4.55}$$

$$\log\left(\frac{\nu_{t,d}}{1-\nu_{t,d}}\right) = \psi_d + \beta_{t,d} \tag{4.56}$$

As the model for the reporting probability is constant, and because this model is largely the same as the model presented in Section 4.5.1, any difference resulting from the addition of an under-reporting mechanism can be summarised by looking at the predictions for x_t and y_t, which are shown in Figure 4.13. Looking at the plot, we now have predictions for the true (unobserved) number of dengue cases over the whole time series. The uncertainty in these predictions is particularly large when in the now-casting and forecasting periods, where uncertainty in the under-reporting, the delayed reporting and in the out-of-sample projection of the disease trend all combine. This suggests that, where data potentially suffer from under-reporting, ignoring it will result in prediction intervals (i.e. those for the predicted total reported counts) which are likely too narrow.

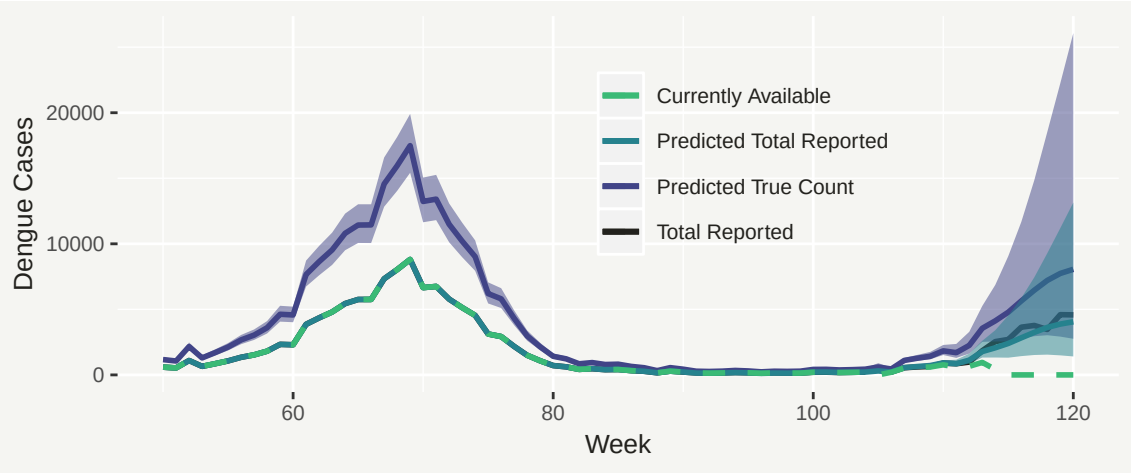

Figure 4.13: Median predicted true (unobserved) dengue cases x_t and total reported cases y_t, with associated 95% intervals. The black line shows the true total reported cases, while the dotted line shows the number of reported cases available to us in the scenario where we are at time $t = 114$.

In this example, the advantage of also taking into account the under-reporting is that public health planners can respond to the estimated number of dengue cases,

the quantity most related to the scale of the public health burden. If the under-reporting is ignored, then the response is instead based on estimates for the total number of cases which will be reported, which may be considerably less, leading to a potentially inadequate response.

4.7 Discussion

In this chapter we have introduced the problem of delayed-reporting and its implications for society. We explained that it is a problem based around prediction, providing a motivation for a statistical approach to the problem. We explored several existing approaches, focusing on (a) approaches based on a Multinomial mixture distribution with either a time stationary Generalized-Dirichlet distribution or a logistic regression and (b) the conditional independence (GLM) approach. Both approaches are very flexible, in terms of incorporating complex spatio-temporal structures. However, we argue that they both have limitations: The approaches based on a Multinomial mixture are not sufficiently flexible to simultaneously capture delay mechanisms which are both heterogeneous in time and over-dispersed, with respect to the Multinomial variance. The GLM approach, on the other hand, does not explicitly model the total counts. This means it relies on capturing the covariance structure of the partial counts well in order to capture the distribution of the total counts well. This is hindered by the assumption that the partial counts are independent, conditional on any covariate or random effects, which in our simulation experiment we found can lead to overestimating the variance of the total counts. To potentially address this, we proposed an extension to this approach (which we refer to as the GLM+) which includes an explicit covariance model for the partial counts, with the aim of better capturing the distribution of the total counts.

We have proposed a general framework based on a Generalized-Dirichlet-Multinomial mixture, where the true total counts are explicitly modelled (unlike the GLM) and where the Multinomial probabilities are a Generalized-Dirichlet whose parameters may vary in space and time. We presented a case study of data on reported dengue fever cases in Rio de Janeiro. In-sample predictive model checking was used to assess the models with respect to how well the distribution of the total number of cases was captured. Out-of-sample predictive checking was also used to assess performance when nowcasting and forecasting. We found that in every test the GDM has the strongest performance, even compared to the GLM+ model which, despite potentially having the most general covariance structure of the three models, was hindered by having an excessively heavy upper tail.

In addition to considering the performance of each model for the particular data set, it is also important to consider other reasons why one might be preferable over the others. The GLM model, for instance, is by far the easiest to implement, espe-

cially in a non-Bayesian setting such as the Generalized Additive Model framework or in an approximate Bayesian setting such as INLA. The GDM, however, lends itself more to a full Bayesian treatment, where Markov Chain Monte Carlo (MCMC) is used, compared to the other frameworks. This is because the effects associated with the true total count and the effects associated with the delay mechanism are separated into different parts of the model and are related to different parts of the data (the total counts and the partial counts, respectively). In the GLM and GLM+ frameworks, meanwhile, all of the effects are in the same model and they end up competing with each other. For this reason, it is possible to obtain a higher effective number of posterior samples per second with the GDM model.

In our view, the GDM framework is the most interpretable of all of the frameworks discussed here. This is because the delay mechanism, and any associated variability, is completely separated from the process which generates total counts. This makes it in turn easier to adapt the model for a given data problem. For example, we can see in Figure 4.12 that the variability in the proportion of cases reported in the first week decreases in the middle of the time series. To capture this, it is a fairly trivial modification to model the logarithm of the dispersion parameters $\phi_{t,s,d}$, as defined in (4.26), using a penalized spline in time. Knowing that variability in the delay mechanism at a certain time is likely to be lower or higher than usual could further improve now-casting precision. Whilst it would be possible to model the Negative-Binomial dispersion parameters θ_d as time-varying in the GLM and GLM+ frameworks, there is no equivalent way of separating temporal structure in the variance of the total counts, from structure in the variance of the delay mechanism, as is possible in the GDM framework.

On the same theme of adaptability, the GDM framework can easily be expanded with an additional module for correcting under-reporting, which may be essential in scenarios where the final observed total count is still a substantial under-representation of the true count. In such situations, allowing for both the delay mechanism and the under-reporting mechanism simultaneously may be crucial for well-informed decision making.

Chapter 5

Conclusion

In this thesis we presented the issue of flawed observation mechanisms and motivated why they should be taken into account, emphasising the risks associated with ignoring them. We evaluated some ways this problem has been previously addressed and argued that the overwhelming majority are too restrictive, generally being bespoke solutions to particular types of flawed observation which are limited in terms of flexibility. For example, we discussed the censored likelihood approach to modelling under-reporting of counts. As this approach relies on prior knowledge of which data are perfectly reported and which are potentially under-reported, its usefulness is limited in cases where all counts are thought to be potentially under-reported.

To address the problem of taking into account flawed observations more generally, we presented a conceptual framework where the true quantity of interest is modelled at a latent level in a Bayesian hierarchical model, with one or more additional layers acting as modules to account for flawed observation mechanisms. In essence, this approach boils down to combining a model for the data we wish we had with one or more models for the mechanisms that relate this to the data we actually have. We argued that this framework has several strengths, namely: the ability to rigorously capture the joint uncertainty in the latent model and the flawed observation mechanisms; the flexibility to be applicable to a wide range of data problems and to allow for complex (e.g. spatio-temporal) structures in both the latent model and in any flawed observation mechanisms; and a high degree of interpretability, due to the way in which the roles of each module are clearly defined. By modelling the flawed observation mechanisms, the framework also provides more precise prediction of the true quantity, compared to methods which do not.

We then spent the rest of the thesis illustrating these points, in the first instance by applying this concept to such a broad range of problems, spanning both the field of health (with the TB and dengue applications), the field of environment (with the tornado and volcano applications) and their interface (with the HAP application). It should be noted that all of the models and results presented in this thesis are one version among hundreds that were tried over the course of this work. It is truly a

tribute to the flexibility of this approach, and more generally Bayesian inference with MCMC, that all of these variations were possible within a single inferential engine. For example, in the application of the delayed-reporting model to dengue, several distinct models for the temporal effects were tried. These included first order random walks, second order random walks and dynamic linear models. Changing between these variations, as well as adjusting prior distributions, amounts to little more than minor modification of the NIMBLE model code. This suggests that, while widely derided for its slow implementation times, the flexibility of MCMC inference allows for rapid innovation in statistical modelling practice. Other modelling frameworks, such as Generalized Linear Models and Generalized Additive Models, can be more restrictive in the range of models they allow. We also note that several of the models we have applied this to have been rather large, in terms of the size of the data but particularly the number of parameters. For example, our model for household air pollution has tens of thousands of parameters.

5.1 Future Research

We have highlighted some potential avenues for future research for the individual flawed observation mechanisms discussed here. For under-reporting, we have suggested further study into how methods such as Bayesian model averaging can alleviate the risk of incorrectly classifying covariates as either belonging to the under-reporting model or the model for the true count. We also assumed in this thesis that covariates related to the model for the true count are distinct from those related to the under-reporting mechanism. In reality, it is reasonable to assume that some covariates may be related to (or proxies for) both. This is the case in Bailey et al. (2005), where a social deprivation indicator is used to both model the true leprosy occurrence rate and to identify under-reported data. In principal there is no reason why it should not be possible to include the same covariate in both parts of the model, but this will introduce further issues of non-identifiability. Generally, we do not see this non-identifiability as an 'evil' which must be exorcised from our model, it merely quantifies our uncertainty arising from the fact that the observed data does not contain information to distinguish between the occurrence rate and under-reporting. However, in some cases, this high degree of uncertainty may limit practicality, and the challenge is then how this non-identifiability can be alleviated. We anticipate that this is achievable through the use of informative priors, or restrictions on the functional shape of covariate terms, or by including data which are known to be completely reported.

For delayed reporting, we have presented a framework (GDM) which we argue is more flexible and more interpretable than its main competitors. Using real data for dengue fever cases, we illustrated how this framework outperforms existing frame-

works across the board (including in now-casting and forecasting precision). Though the difference in performance was not huge for this data, we would nonetheless opt for the GDM for its interpretability, and for the option of modelling the variance of the delay mechanism as non-stationary. However, we believe there is still room for improvement. In particular, the specification of the mean delay structure in terms of the expected relative proportion reported after each interval is not very intuitive. It also limits possibilities for reducing the complexity of the model where the delay structure is more simple. We have since developed an alternative formulation for the mean delay structure, which begins with a model for the expected cumulative proportion of cases reported after each delay interval. This can then be converted into relative proportions, so that the rest of the model is the same as before. We have found this model to be equivalently flexible to the one presented here, but we believe it is more intuitive and it allows for simpler mean delay structures.

We also presented a way of taking into account under-reporting in the final observed count and extended our model for the dengue fever data to illustrate a simple version of this approach. We believe this framework in particular should be further explored, though this may require additional data in the form of covariates to inform the under-reporting mechanism, or an exploration of how the under-reporting model may be linked to the delayed reporting model.

More broadly, whilst we introduced the idea that in some situations different flawed observation mechanisms may interact, and suggested how our framework could accommodate this, future research should include an investigation into the effectiveness of this solution.

5.2 Final Remarks

The types of flawed observation mechanisms discussed here are only a few among many others which this thesis does not cover. For example, under-reporting is just one issue count data may suffer from. Counts can also be over-reported or, more generally, misreported. Moving beyond count data, the literature on left-censoring, right-censoring, interval censoring and truncation in the field of survival analysis is vast, and there is also well-established work on modelling data with missing values (including where they are missing in a structured way). There may exist approaches to modelling these mechanisms which fall into the mindset discussed here but, where they don't, there is certainly scope for applying this powerful framework to these problems.

The main contribution of this thesis is to highlight the power of the framework: we were able provide predictive inference for true counts of disease cases and natural hazards, where all of the available data were potentially under-reported; we were able to predict the use of 8 key fuels for cooking, in the context of incomplete

surveys and substantial sampling biases; and we were able to predict (now-cast) unobserved disease cases based on available partial counts and even future disease trends for which we have no information. Notably, it was only by taking into account the various flawed observation mechanisms that a combined model for the use of individual fuels for cooking was made possible. This model has since been adopted by the WHO for estimating and forecasting the proportion of people using polluting fuels for cooking. These estimates have played a key role in official publications, such as the 2019 'energy progress report' (IEA, IRENA, UNSD, WB, WHO, 2019).

To conclude, we feel we have clearly demonstrated that this approach to taking into account flawed data should be advocated in future modelling practice, but particularly in the fields of environment and health.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. Wiley.

Aitchison, J. and C. H. Ho (1989, 12). The multivariate Poisson-log normal distribution. *Biometrika 76*(4), 643–653.

Atlas (2013). The brazilian municipal human development index. Technical report.

Bailey, T., M. Carvalho, T. Lapa, W. Souza, and M. Brewer (2005). Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology 15*(5), 335 – 343.

Barbosa, M. T. S. and C. J. Struchiner (2002, 02). The estimated magnitude of AIDS in Brazil: a delay correction applied to cases with lost dates. *Cadernos de Saude Publica 18*, 279 – 285.

Bastos, L., T. Economou, G. M., V. D., T. Bailey, and C. Codeço (2017, sep). Modelling reporting delays for disease surveillance data. `https://arxiv.org/abs/1709.09150`. Accessed: 2019-02-14.

Besag, J., J. York, and A. Mollié (1991, Mar). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*(1), 1–20.

Bonjour, S., H. Adair-Rohani, J. Wolf, N. G. Bruce, S. Mehta, A. Prüss-Ustün, M. Lahiff, E. A. Rehfuess, V. Mishra, and K. R. Smith (2013). Solid fuel use for household cooking: country and regional estimates for 1980–2010. *Environmental health perspectives 121*(7), 784.

Broemeling, L. (2013). *Bayesian Methods in Epidemiology*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Clarke, P. and R. Hardy (2007). *Methods for handling missing data. In Epidemiological Methods in Life Course Research*. Oxford University Press.

Crainiceanu, C., D. Ruppert, and M. Wand (2005). Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software, Articles 14*(14), 1–24.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik (2017). Programming with models: Writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics 26*(2), 403–413.

Dobson, A. and A. Barnett (2018). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Dvorzak, M. and H. Wagner (2016). Sparse bayesian modelling of underreported count data. *Statistical Modelling 16*(1), 24–46.

Economou, T., D. B. Stephenson, and C. A. T. Ferro (2014, 12). Spatio-temporal modelling of extreme storms. *Ann. Appl. Stat. 8*(4), 2223–2246.

Elsner, J. B., T. Fricker, H. M. Widen, C. M. Castillo, J. Humphreys, J. Jung, S. Rahman, A. Richard, T. H. Jagger, T. Bhatrasataponkul, C. Gredzens, and P. G. Dixon (2016). The relationship between elevation roughness and tornado activity: A spatial statistical model fit to data from the central great plains. *Journal of Applied Meteorology and Climatology 55*(4), 849–859.

England, P. and R. Verrall (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal 8*(3), 443–518.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2014, November). *Bayesian Data Analysis, Third Edition (Chapman and Hall/CRC Texts in Statistical Science)* (Third ed.). London: Chapman and Hall/CRC.

Greer, B. A., J. D. Stamey, and D. M. Young (2011). Bayesian interval estimation for the difference of two independent poisson rates using data subject to underreporting. *Statistica Neerlandica 65*(3), 259–274.

Höhle, M. and M. an der Heiden (2014, 6). Bayesian nowcasting during the stec o104:h4 outbreak in Germany, 2011. *Biometrics 70*(4), 993–1002.

Ibrahim, J., M. Chen, and D. Sinha (2001). *Bayesian survival analysis*. Springer series in statistics. Springer.

IEA, IRENA, UNSD, WB, WHO (2019). Tracking sdg 7: The energy progress report 2019. `https://www.who.int/airpollution/data/household-energy-database/en/`.

Kennedy, W. and J. Gentle (1980). *Statistical Computing*. Marcel Dekker.

Kirk, P. J. (2014). An updated tornado climatology for the UK: 1981–2010. *Weather 69*(7), 171–175.

Lawson, A. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Third Edition*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software, Articles 63*(19), 1–25.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best (2009). The bugs project: Evolution, critique and future directions. *Statistics in Medicine 28*(25), 3049–3067.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin 23*(2), 213–225.

Morales, I., H. Salje, S. Saha, and E. S. Gurley (2016). Seasonal distribution and climatic correlates of dengue disease in Dhaka, Bangladesh. *The American Journal of Tropical Medicine and Hygiene 94*(6), 1359–1361.

Moreno, E. and J. Girón (1998). Estimating with incomplete count data a bayesian approach. *Journal of Statistical Planning and Inference 66*(1), 147 – 159.

Oliveira, G. L., R. H. Loschi, and R. M. Assunção (2017). A random-censoring poisson model for underreported data. *Statistics in Medicine 36*(30), 4873–4892.

Papadopoulos, G. and J. M. C. S. Silva (2008). Identification issues in models for underreported counts. *Discussion Paper Series, Department of Economics, University of Essex* (657).

Papadopoulos, G. and J. S. Silva (2012). Identification issues in some double-index models for non-negative data. *Economics Letters 117*(1), 365 – 367.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rehfuess, E., S. Mehta, and A. Prüss-Usün (2006). Assessing household solid fuel use: multiple implications for the millennium development goals. *Environmental health perspectives 3*(114), 373–378.

Renshaw, A. E. and R. J. Verrall (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal 4*(4), 903–923.

Rougier, J., S. Sparks, K. Cashman, and S. Brown (2018, 1). The global magnitude-frequency relationship for large explosive volcanic eruptions. *Earth and Planetary Science Letters 482*, 621–629.

Salmon, M., D. Schumacher, K. Stark, and M. Höhle (2015). Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal 57*(6), 1051–1067.

Shaddick, G., M. L. Thomas, A. Green, M. Brauer, A. Donkelaar, R. Burnett, H. H. Chang, A. Cohen, R. V. Dingenen, C. Dora, S. Gumy, Y. Liu, R. Martin, L. A. Waller, J. West, J. V. Zidek, and A. Prüss-Ustün (2017). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 67*(1), 231–253.

Shaddick, G. and J. Zidek (2015, 6). *Spatio-Temporal Methods in Environmental Epidemiology*. CRC Texts in Statistical Science. United Kingdom: Chapman & Hall.

Shaweno, D., J. M. Trauer, J. T. Denholm, and E. S. McBryde (2017, Oct). A novel bayesian geospatial method for estimating tuberculosis incidence reveals many missed TB cases in Ethiopia. *BMC Infectious Diseases 17*(1), 662.

Silva, M. M. O., M. M. de Souza Rodrigues, I. A. D. Paploski, M. Kikuti, A. M. Kasper, J. S. Cruz, T. L. Queiroz, A. S. Tavares, P. M. Santana, J. M. G. Araújo, A. I. Ko, M. G. Reis, and G. S. Ribeiro (2016). Accuracy of dengue reporting by national surveillance system, Brazil. *Emerging infectious diseases 22 2*, 336–9.

Stamey, J. D., D. M. Young, and D. Boese (2006). A bayesian hierarchical model for poisson rate and reporting-probability inference using double sampling. *Australian and New Zealand Journal of Statistics 48*(2), 201–212.

Stoner, O. (2018). Correcting under-reporting in historical volcano data. *Proceedings of the 33rd International Workshop on Statistical Modelling 1*, 288–292.

Stoner, O. and T. Economou (2019). Multivariate hierarchical frameworks for modelling delayed reporting in count data. `https://arxiv.org/abs/1904.03397`. Accessed: 2019-04-22.

Stoner, O., T. Economou, and G. Drummond Marques da Silva (2019a). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*.

Stoner, O., G. Shaddick, T. Economou, S. Gumy, J. Lewis, I. Lucio, and H. Adair-Rohani (2019b). Estimating household air pollution: A multivariate hierarchical

model for the use of polluting fuels for cooking. `https://arxiv.org/abs/1901.02791`. Accessed: 2019-02-18.

Stoner, O., G. Shaddick, T. Economou, S. Gumy, J. Lewis, I. Lucio, G. Ruggeri, and H. Adair-Rohani (2019c). Multivariate hierarchical modelling of household air pollution. *Proceedings of the 34th International Workshop on Statistical Modelling 2*, 242–247.

Tibbits, M. M., C. Groendyke, M. Haran, and J. C. Liechty (2014). Automated factor slice sampling. *Journal of Computational and Graphical Statistics 23*(2), 543–563.

United Nations (2018). World urbanization prospects: The 2018 revision. Technical report.

Wang, X., M. Zhou, J. Jia, Z. Geng, and G. Xiao (2018). A bayesian approach to real-time monitoring and forecasting of chinese foodborne diseases. *International Journal of Environmental Research and Public Health 15*(8).

Winkelmann, R. (1996, Dec). Markov chain monte carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics 21*(4), 575–587.

Winkelmann, R. (1998). Count data models with selectivity. *Econometric Reviews 17*(4), 339–359.

Winkelmann, R. (2008). *Econometric Analysis of Count Data* (5th ed.). Springer Publishing Company, Incorporated.

Winkelmann, R. and K. F. Zimmermann (1993). Poisson-logistic regression. *Discussion Papers, Department of Economics, University of Munich 93*(18).

Wong, T.-T. (1998). Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation 97*(2), 165 – 181.

Wood, S. (2016). Just another gibbs additive modeler: Interfacing jags and mgcv. *Journal of Statistical Software, Articles 75*(7), 1–15.

Wood, S. N. (2017, May). *Generalized Additive Models: An Introduction with R, Second Edition (Chapman and Hall/CRC Texts in Statistical Science)* (Second ed.). London: Chapman and Hall/CRC.

World Health Organization (2012). *Assessing tuberculosis under-reporting through inventory studies*. Geneva, Switzerland.

World Health Organization (2014). *WHO guidelines for indoor air quality: household fuel combustion*.

World Health Organization (2016). *Global Tuberculosis Report*. Geneva, Switzerland.

World Health Organization (2018a). Household energy database. `https://www.who.int/airpollution/data/household-energy-database/en/`. Accessed: 2018-12-17.

World Health Organization (2018b, September). WHO dengue and severe dengue fact sheet. `https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue`. Accessed: 2019-02-14.

World Health Organization (2018c). WHO press release. `https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action`. Accessed: 2018-12-17.